

Clean, green conferencing

Big conferences are good for science. But because many researchers fly in, these events are also bad for the environment. What can be done to redress the balance?

Consider this contradiction. Compared with most other professions, scientists are probably better informed and more concerned about climate change. Yet they also fly more than most, generating significant greenhouse-gas emissions. Last month, for example, 31,000 neuroscientists descended upon San Diego for their annual meeting. Even many of those based in the United States flew in.

Researchers should consider what to do about this, because politicians are unlikely to take any action. In countries that have signed the Kyoto Protocol, companies are starting to cut emissions through carbon-trading schemes. But airlines will not join the party, because aviation exhaust gases will lie outside the Kyoto Protocol until at least 2012. That's a problem: the sector generates 3.5% of global emissions, and its contribution is expected to double in the next 15 years.

What should scientists do? Some advice is as well worn as it is tough to take: think about going by train, bike or camel, or whether to go at all. These are not palatable ideas. Conferences are often too distant to reach, except by plane, but good science depends on the exchange of ideas. Nonetheless, it would be worth research groups considering whether they can send fewer members to conferences, and whether more distant and less important meetings can be missed altogether.

One alternative to missing events is to tot up the total emissions incurred by a flight, and invest in small-scale projects to cancel out the emissions. This could involve cutting carbon from other sources, by paying for solar panels, for example, or giving villagers in the developing world stoves that burn more cleanly. The web is teeming

with companies that will calculate your next trip's emissions, work out how much it will cost to cut these emissions from other sources (see News Feature, page 268) and invest that amount in selected projects. Bingo — clean, green conferencing.

Some research organizations, such as the Tyndall Centre for Climate Change Research in Norwich, UK, already do this on a routine basis without incurring significant costs. Other institutes should consider following suit. Alternatively, academic societies might choose to add the costs to the meeting fees.

But before doing so, take a good look at the companies that offer to do this 'carbon offsetting'. The problem is verification. Under Kyoto, companies can invest in well regulated mitigation projects, such as schemes to collect methane from landfill sites. Until offsetting projects are regulated in the same way, there is no guarantee that the firms involved have done their homework. Many offer to plant trees, for example, but forestry is not an accepted emissions-management strategy under Kyoto, partly because it offers no long-term guarantee of soaking up carbon. Trees can be cut down or burnt, especially if local people need them for fuel or economic gain.

To avoid such pitfalls, ask offsetting companies a few questions before investing. How transparent are they? Do they, for example, ask independent scientists to scrutinize their projects? And do they take a truly international outlook? There is little point in restricting investment to schemes in one particular country, as some companies do. Before you invite your colleagues to jet over to your own carbon-neutral conference, make sure that it actually is. ■

A chance for growth

With the right safeguards, a national institute could give a much-needed boost to agricultural research in the United States.

Plans are afoot to create more fertile pastures for research within the US Department of Agriculture (USDA). A proposed National Institute for Food and Agriculture, which could eventually have as much as \$1 billion a year to spend, would be associated with the USDA, but operate independently of it.

At first sight, this seems like a good opportunity to reinvigorate US agricultural research. Science at the USDA has been hamstrung for years, not by a lack of funding but by structural problems. Only about 15% of its \$2-billion annual research spend goes to competitive grants; the figure at the National Institutes of Health (NIH) is 70%. Some researchers are wary of the USDA's research programmes, fearing that they lack a long-term strategy. Farmers, consumers, researchers and agribusiness would all benefit if more of the cash went on competitive, peer-reviewed grants.

Advocates of the proposed institute, which is likely to be considered by Congress this month, say it could do this by operating along the same lines as the NIH, which has a long tradition of distributing grants on the basis of merit, without interference from its parent, the Department of Health and Human Services, or from the Congress.

But matters are unlikely to be so straightforward at the USDA, in part because agribusiness is accustomed to exercising a strong influ-

ence over the department. It has championed the institute's creation and may expect to retain influence over its research programmes.

To constrain that influence and ensure the institute's independence, some minimum conditions must be met: its director should be a reputable and independent scientist, for example, and so should the members of its advisory panel. These people will put their reputations on the line when they take up their posts, and that should ensure they form a bulwark against commercial interference.

Harder to guarantee will be an understanding on the part of the congressional committees that fund the USDA that the new institute is to be left alone. In the past, the USDA's competitive, peer-reviewed research programmes have been eroded by a culture that demands that each politician receive something for the state they represent.

The committees should consider what the institute stands to lose if they do not take a hands-off approach. The institute is expected to tackle contentious questions on food, its relation to human health, and environmental issues. These questions can only be answered effectively by an institute that has the trust of the public. Equally importantly, the answers will only be accepted internationally if the institute is seen to take on the toughest questions and report the results honestly, irrespective of the corporate economic impact. ■





German marks
Rating system set to see how institutes measure up
p260



Worried sick
Health agency injects urgency into hunt for pandemic flu vaccine
p261



Benchmark
Molecular-biology lab celebrates its 30th anniversary
p262



Screen breaks
Italians take time out to protest about university reforms
p264

Campaign to fight malaria hit by surge in demand for medicine

David Cyranoski, Tokyo

An alarming shortfall of a key drug is undermining an international drive to reduce the malaria death toll. A rise in demand has led to a shortage of artemisinin, the main treatment for malaria that is resistant to conventional therapies, the World Health Organization (WHO) announced on 8 November.

Artemisinin is extracted from the wormwood plant, *Artemisia annua*, which grows wild in southern China and Vietnam. Combined with other drugs, its derivatives, such as artesunate and artemether, can clear symptoms of malaria in three days. Malaria currently kills about a million people every year, mainly in Africa.

In 2001, the WHO recommended that artemisinin-based combination therapies, or ACTs, should be used in countries where there is resistance to drugs such as chloroquine. Artemisinin-based drugs are more expensive than conventional treatments, in part because large doses are required.

The WHO reached a high-profile agreement with drug firm Novartis in 2001 for the company to supply one such ACT — artemether–lumefantrine (Coartem) — at cost price. In May this year, the Geneva-based Global Fund to Fight AIDS, Tuberculosis and Malaria took the further step of requiring all of its malaria grants to be used on artemisinin-based drugs.

Production problems

Forty countries have started using ACTs, 18 of them since January. Requests for Coartem treatment courses alone are projected to rise to 10 million in 2004 — up from 220,000 in 2001 — and to hit 60 million next year.

The challenge of meeting this increasing demand took a turn for the worse in July, when the major supplier of artemether, Kunming Pharmaceutical of Yunnan, China, said it could not keep up. In October, the WHO learned that Novartis would not be able to deliver much more than half of the 4.5 million doses it requested by March 2005.



Natural high: demand for products of the wormwood plant, which is used to make antimalarial drugs, has soared.

“We were shocked — nobody thought the situation would be this bad,” says Andrea Bosman, medical officer for the WHO’s Roll Back Malaria campaign. Bosman says she is expecting about 2.4 million doses from Novartis in this period. “Are other companies better off? Probably not.”

Countries such as Ethiopia, where up to 70,000 people died from malaria last year, will be hit particularly hard. It will only receive about half of the two million doses that its government has requested.

Cultivation and extraction take a minimum of nine months, and Li Su, a spokeswoman for Kunming Pharmaceutical, says the company was not given enough notice to

prepare. “Before June 2004, we never got a guarantee of a large order from Novartis,” she says. The company is now ramping up production. A factory that can produce 20 tonnes of artemether a year — the full amount requested by Novartis — is due to open next autumn.

But manufacturing capacity only matters if the raw material is available. Increasing demand has pushed up the price and forced producers to use low-quality, low-yield leaves. Kunming Pharmaceutical will open its own plantation in February.

With production looking bleak for months to come, Bosman says that one of the greatest concerns is the further spread of ineffective fake ACTs. “This will feed a huge black market,” she says.

Work is also under way to find varieties that grow well in local climates in Africa. Tanzania already has a promising variety and should be able to provide 20 tonnes by 2006, says Bosman.

Chemical fix

Long-term, hopes are resting on the development of synthetic artemisinin-based drugs, which avoid the unreliability of cultivation. Scientists at the University of California, Berkeley, have created transgenic bacteria that can make a precursor to artemisinin. (V. J. J. Martin *et al. Nature Biotechnol.* **21**, 796–802; 2003). They are now trying to make bacteria that can recreate the entire artemisinin synthetic pathway, aiming for industrial production by 2009. They say this would slash the price to one-tenth of its current level.

Another synthetic solution may be closer to hand. A chemically synthesized compound called RBx-11160, which mimics the chemical and biological properties of artemisinin (J. L. Vennerstrom *et al. Nature* **430**, 900–904; 2004), is already in clinical trials in Britain.

Novartis says it remains optimistic of getting back on track. But the WHO, after years of trying to persuade countries to use ACTs, has been forced to recommend that they use quinine until stocks recover. ■

Europe's stem-cell workers pull together

Federica Castellani, Munich

As California gears up to become a world force in research on human embryonic stem cells, European scientists are banding together, buoyed by recent changes in Spanish and French law.

The European Stem Cell Network held its inaugural meeting in Seville, Spain, on 12 November. It brought together scientists from 14 European nations and Israel in a bid to promote collaboration.

"We clearly need a network in Europe," says meeting organizer Bernat Soria, director of the Institute of Bioengineering at Miguel Hernández University in Alicante. "Stem-cell research is in its infancy — we want to collaborate and cooperate, not compete with each other, just now."

"In Europe we will never have the same resources as California will have," says Outi Hovatta, a stem-cell researcher at the Karolinska Institute in Stockholm. "But we are trying to reinforce ourselves by

combining resources and experiences."

As if to set an example, Soria himself cemented a collaboration at the meeting. His group's diabetes research will be bolstered by new stem-cell lines from the Karolinska Institute.

Europe has been deeply divided over stem-cell research, but an increasing number of countries are permitting the work. Last month both Spain and France joined countries such as Britain and Sweden in passing laws to allow research on human embryonic stem cells.

Other nations remain opposed, however. German scientists, for example, much like their US colleagues outside California, can work only with first-generation stem-cell lines created before 1 January 2002. These older stem-cell lines are of limited use in front-line research, scientists complain.

"The legislative situation in Europe is very complicated," says Oliver Brüstle, a stem-cell researcher at the University of

Bonn. "But if we want to work internationally and exchange data we really need to be able to work with the same cell lines."

Scientists in the field agree that they need a single voice to speak to governments and put pressure on decision-makers for a common legislation. The new network, which will meet again in Edinburgh next March, could provide such a voice, says Soria.

And the need for common legislation is likely to become more acute — the European Commission's first explicit call for proposals for research involving human embryonic stem cells has just closed and funds could be forthcoming.

To secure funding from the commission, research projects have to involve multinational groups. But what the legal situation would be for a German scientist taking part in a project involving human embryonic stem cells has yet to be tested, Brüstle says. He adds that he will not work on any cell lines that are illegal in Germany. ■

Grade expectations for German research institutes

Quirin Schiermeier, Munich

To help students and researchers separate the wheat from the chaff, Germany's science council has proposed a rating system for the country's publicly funded research institutes.

At a meeting in Hamburg last week, the science council recommended that departments in up to 50 disciplines be regularly marked by independent experts. All German research organizations, including the Max Planck Society, have agreed to participate.

When fully established, probably in 2006, the ratings will be an easy way of comparing the strengths of different labs, says the science council.

Although the details have yet to be hashed out, grades are expected to be based on a seven-point scale and awarded once every five or six years. A department will receive the highest mark only if more than half of its research activities are considered to be of top international quality; the lowest mark will go to departments whose output falls below national standards. The science council is encouraging the development of both hard and soft indicators to be applied in varying degrees depending on the field, which should spare anthropologists and quantum physicists from being measured on the same criteria.

The ratings are likely to affect future funding decisions, but they are not meant to become the decisive factor. This will differentiate the system from one that



Solid foundations: German universities, such as Heidelberg, are squaring up to a future rating system.

currently exists in Britain — the Research Assessment Exercise, which uses grading to assign some government grants.

The British audits are thought to have improved the competitiveness of UK research, but critics say they are excessively expensive and time-consuming, and tend to overemphasize quantitative indicators (see *Nature* 418, 6; 2002).

Similar objections are bubbling up in Germany. "Rankings are very popular among politicians and the public, but they are never objective, and sometimes they are even detrimental to science," says Wolfgang Baumeister, director of molecular structural biology at the Max Planck Institute for

Biochemistry in Martinsried, near Munich. "They foster mainstream research and the hunt for short-term accomplishments instead of true innovation."

The proposal is being backed by the DFG, Germany's main funding agency. An institute will be created in Bonn to develop and test rating methodologies, and a pilot project designed to score sociology and informatics departments is scheduled to start in 2005. ■

For more news and analysis go to
news @ nature.com
www.nature.com/news



Ill-prepared: although flu vaccines are made every year, the world is not equipped to fight a pandemic.

WHO calls for vaccine boost to prepare for flu pandemic

Erika Check, Washington

Alarmed by bird flu's grip on southeast Asia, the World Health Organization (WHO) is calling for action. Last week, WHO officials met with vaccine makers, public-health experts and government representatives in a bid to speed up the production of flu vaccines to avert a global pandemic.

The three flu pandemics of last century — in 1918, 1957 and 1968 — were sparked when avian flu jumped the species barrier and became infectious in humans. The WHO has been warning for two years that conditions are right for a new pandemic — yet vaccines are still not available.

Theoretically, little stands in the way of preparing pandemic flu vaccines — the annual manufacture of vaccines for normal flu is routine. But companies and governments have been reluctant to invest in a vaccine that would never be used if the threat of a pandemic abates.

"The vaccines need to be tested, and that costs money," said Klaus Stöhr, coordinator of the WHO's global influenza programme, at the 12 November meeting. "That is the single most important barrier to pandemic vaccine development."

The strain of bird flu now circulating in Asia — known as H5N1 — has already killed 44 people this year, but does not seem able to spread from person to person, yet. WHO officials said at the meeting that the H5N1 strain seems to be increasing in virulence in chickens and mice, and may have found a reservoir in domestic ducks. The ducks can be infected for a lengthy period while appearing healthy, and could spread the disease back into poultry, and possibly humans, said WHO officials. The incubation of the virus in these different species may favour its evolution

into a strain that can spread among humans.

But just two companies are currently testing a pandemic flu vaccine, and only a handful of countries have funded clinical trials of the vaccines. The WHO urged an acceleration of these efforts. "We can get our homework done now to ensure that when it matters most to get vaccines produced, it can happen immediately," Stöhr said. "We don't want to miss this chance."

Stöhr also noted the need to resolve inconsistencies in policies among international regulatory agencies, to avoid duplicating costly clinical trials (see *Nature* 432, 137; 2004).

But the initiative still has challenges to face, said Stöhr. A major barrier is that patents on a crucial technology, called reverse genetics, are held by one company, MedImmune in Gaithersburg, Maryland. The technology has the potential to accelerate the process by which vaccines are selected and produced. Other companies hold associated patents. Representatives from industry told the meeting that dealing with patent issues has proved to be a major impediment for vaccine development in the past.

Some governments have begun to take action. The US National Institutes of Health is funding pandemic flu vaccine trials. The US health department also said on 9 November that it has issued a \$10-million contract to secure a year-round supply of eggs needed to grow the volume of virus necessary for a pandemic flu vaccine. And Japan has expressed interest in funding trials, say WHO officials.

These efforts are admirable, but they are not enough, said Arlene King of Canada's Public Health Agency, who attended the Geneva meeting. "Pandemic influenza will be the biggest public-health emergency we ever face," she said. ■

Grant-transfer plan paves the way for European mobility

Barbara Simm, Munich

European researchers look set to enjoy a mobile future. The scientific borders of ten nations have been loosened to allow national grants to be spent in other countries.

The requirement for grants to be spent on research in the home nation was seen by many as a hindrance to mobility in Europe. And that in turn was cited as a major competitive disadvantage compared with the United States.

Now the heads of Europe's national research councils (collectively known as EUROHORCS) have hammered out an agreement to let researchers take their grants to an institute in a participating country without needing approval. The transfer can be made as long as the funded project is already under way, and its funding has at least six months left to run.

The agreement was signed on 22 October by the heads of 12 research councils from ten European nations, but it was made public only last week. Among those who signed were Britain's Engineering and Physical Sciences Research Council (EPSRC) and its Particle Physics and Astronomy Research Council. The main national research agencies in France, Switzerland and Finland intend to join by the end of the year, and the remaining UK research councils are likely to follow suit, says Christoph Mühlberg, head of international affairs at Germany's research funding body the DFG, which currently chairs EUROHORCS.

The DFG already has a number of bilateral agreements for cross-border grant validity, for example with the EPSRC and the Austrian Science Fund (FWF). But Mühlberg says that the EUROHORCS agreement is a major step towards border-free research mobility throughout Europe.

Many researchers are enthusiastic about the agreement. "Being able to move without a lot of red tape, grant-wise, makes it infinitely easier to establish a scientific career abroad," says Marcus Koch, an Austrian plant scientist who moved to the University of Heidelberg in Germany last year, shortly after the bilateral agreement was signed between the DFG and the FWF. "Bringing my Austrian grant money to Germany spared me so much time and trouble."

Other countries participating in the EUROHORCS programme include Spain, Austria, Belgium and the Netherlands. ■

Molecular biology enjoys double celebration

Markus Wagner

The club of life scientists launched by nuclear physicist-turned-biologist Leo Szilard this week celebrates the fortieth anniversary of its transformation into EMBO, the European Molecular Biology Organization.

At the same time it celebrates the thirtieth anniversary of EMBO's most visible manifestation — the European Molecular Biology Laboratory (EMBL).

The celebrations are not ostentatious — a glass of champagne, a symposium. EMBO and EMBL, both based in Heidelberg, Germany, have gained status in the past ten years, but the financial temperature has always been chilly. At a time when the needs of molecular biology have grown, particularly for computational support to make sense of genomic data, funding has followed only grudgingly.

Despite such reluctance, EMBO, which runs a prestigious club of EMBO fellows as well as select meetings, has increased its modest budget by half in the past decade, allowing it to broaden its scope. This week it announced plans to extend its activities in central eastern Europe in the form of 'installation grants' to help young scientists set up their first lab. To begin with, this initiative is being financed by the Howard Hughes Medical Institute.



Over the years: the European Molecular Biology Laboratory has become one of the most cited biomolecular institutions.

EMBL, with its extensive infrastructure and 1,300 staff based at Heidelberg and four outstations in different countries, has always been a more expensive proposition. Over the years, it has fought, and mostly won, hard battles with the governments of the 17 member states that finance it. Its budget has more than quadrupled since 1981. The fights have often been bitter — in the early 1990s, Italy, a major funder, threatened to pull out if EMBL did not create special labs in Italy; the lab complied. But EMBL is now the most cited molecular biology research institution outside the United States.

Fleeing the Cuban missile crisis in 1962, Szilard found refuge at CERN, the European particle-physics laboratory in Geneva, which celebrated its fiftieth anniversary this year



(see *Nature* 430, 824–827; 2004). CERN had been created to counteract the drain of nuclear scientists to the United States, and to promote international cohesion in post-war Europe.

Szilard envisioned that EMBO would do the same for biologists. But many critics argued that funding of a bricks-and-mortar European institute would only take money away from national projects. They complained that while CERN's huge particle accelerators could not be afforded by one country alone, molecular biology was cheap and could be done at home. It is no longer cheap — and its basic economic value is no longer in doubt. Critics are quieter these days. ■

Stalemate over fusion project threatens to provoke split

Jim Giles

The European Union (EU) is considering going it alone with plans to build ITER, an international attempt to develop fusion as an energy source.

Research ministers from EU member states will discuss plans to build the €4.7-billion (US\$6.1-billion) device without the help of international partners when they meet in Brussels on 25 and 26 November. Officials at the European Commission, the EU's executive arm, stress that this is a fall-back position should arguments over where to site the reactor not be resolved.

Hopes for a resolution are not high. France and Japan have both proposed sites for the reactor, which will attempt to create fusion energy by heating a plasma constrained in a magnetic field. The proposals are considered to have equal merit, and parties have been deadlocked for more than a year. European nations want the

reactor to be based at Cadarache in France; the United States and South Korea favour the Japanese site at Rokkasho.

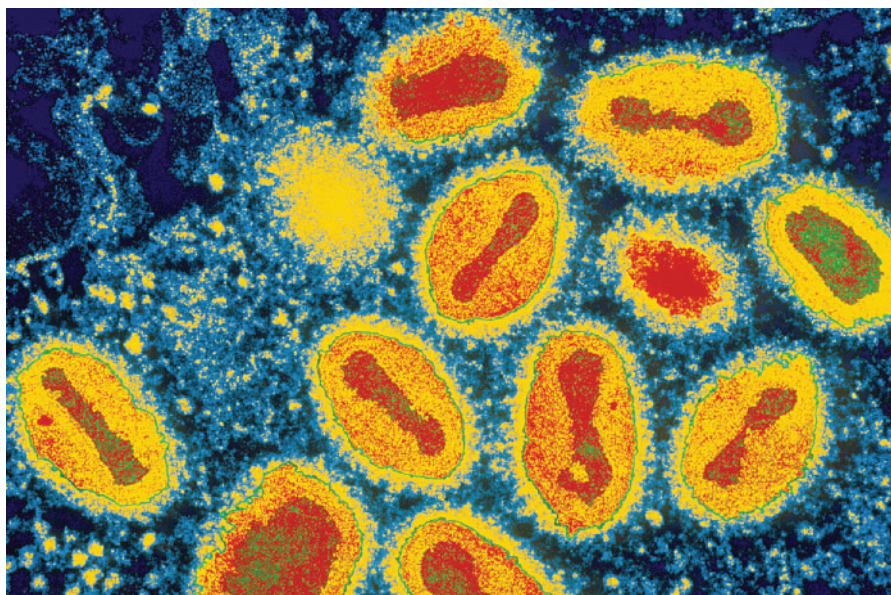
The latest international meeting, held on 8 and 9 November in Vienna, ended yet again without decision. The dispute attracted attention when the Reuters news agency reported that an EU official had said that Japan was going to back the French site. Japanese officials angrily denied the story and, according to a European source, stiffened their resolve not to back down.

EU ministers will now consider an analysis of their ability to host the project without Japan, South Korea and the United States. That trio are currently offering to provide 30–40% of the construction and launch costs, says Fabio Fabbi, commission spokesman for research. He says the commission will consider asking for extra funding from France, Russia and China, or find cheaper ways of building the reactor.

"For both financial and technological reasons, I think that would be very difficult," says Takahiro Hayashi, deputy director of Japan's office of fusion research. "This has always been a project based on international cooperation. Giving up on that would be deplorable."

EU ministers may decide to continue negotiations. Details of the Vienna talks have not been revealed, but it is believed that Japan was offered a deal under which it would end its bid to be host in return for a greater role in other research projects associated with ITER.

Those involved in the decision-making all say they would prefer a full international programme, but stress that talks cannot continue indefinitely. One official, who asked not to be named, points out that ITER currently has backing from the highest political level of the union. "But in another year nobody can say what the level of support will be," the official adds. ■



Varying variola: researchers may soon be allowed to add a 'marker gene' to the smallpox virus.

Britain to combat conflicts of interest in drug regulators

Jim Giles, London

Tough measures designed to excise conflicts of interest from committees that regulate drugs have been proposed by the British government.

If the proposals are accepted, advisers on the panels that assess drug safety and performance, and report to the Medicines and Healthcare products Regulatory Agency (MHRA), will have to relinquish all financial interests in the pharmaceutical industry, health minister Norman Warner said on 11 November.

Committee members would also have to declare financial benefits, such as conference costs paid for by industry.

The move comes as pressure mounts on pharmaceutical companies and regulators in Europe and the United States. The MHRA has been the subject of television and newspaper investigations in recent months, which have alleged numerous conflicts of interest on the part of staff and advisers.

In the United States, officials at the Food and Drug Administration (FDA) have been accused of mishandling health scares associated with the painkiller Vioxx (rofecoxib) and a widely used class of antidepressants (see *Nature* 431, 122–124; 2004). In both cases, critics accused the FDA of acting in the interests of industry rather than patients.

The UK proposals, which are open for consultation until February, meet many of the demands made by patients' rights activists in the wake of these stories. In addition to ruling out direct pharmaceutical interests, members would have to declare relevant interests held by family members. Scientists on the committee would also have to say whether they have done any research relating to a particular product, even if that work was not funded by a pharmaceutical company.

Patients' groups in the United States have welcomed the proposals. "The British regulators are moving in the right direction and opening up the secretive club that binds drug-industry representatives and regulators," says Vera Hassner Sharav of the Alliance for Human Research Protection in New York.

The Association of the British Pharmaceutical Industry insists that the MHRA committees are already impartial, because members do not take part in discussions on drugs in which they have a financial interest. But the association accepts that "justice must not only be done but be seen to be done".

Unanimous vote approves tweak to smallpox genome

Erika Check, Washington

An influential committee at the World Health Organization (WHO) has voted in favour of modifying genes in the smallpox virus, and in the vaccine strains that eradicated the disease.

If enacted, the recommendations would overturn decades of policy that govern the two remaining samples of the variola virus, which causes smallpox. The samples are kept at the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, and at a lab run by the Russian government in Novosibirsk, Siberia.

Scientists have traditionally been wary of allowing these samples to be genetically modified. But on 5 November, the WHO Advisory Committee on Variola Virus Research voted unanimously to permit one specific genetic manipulation of the virus, says WHO spokeswoman Maria Cheng. The members also voted to permit genetic modification of the vaccine strain that eradicated the disease from the globe.

The advisers said that the moves would speed research on smallpox treatments. The committee proposes that researchers be allowed to insert a 'marker gene' into the variola genome, which would make the virus glow green. This would allow scientists to detect quickly whether a potential treatment has killed the virus.

The recommendations will be considered early next year by the WHO's director-general and executive board, and possibly later by the 192-member World Health Assembly.

"The use of a marker gene would permit greater rapidity for things such as screening antivirals, and that in turn would decrease the contact time that researchers have with the virus," says Inger Damon, chief of the

CDC's poxvirus section, who attended the WHO meeting.

The only potential drug treatment for smallpox, an antiviral compound called cidofovir, must be given intravenously, although scientists are trying to develop an oral version. Ideally, researchers would like to have more than one treatment option in case drug-resistant smallpox strains emerge naturally or are created by deliberate engineering. The current vaccines against smallpox are also imperfect as they are either inefficient or have potentially life-threatening side-effects.

But the WHO proposals have met with fierce criticism. Jonathan Tucker, a non-proliferation analyst at the Monterey Institute of International Studies in California, says that allowing the modification could give some countries the idea that the United States and Russia are trying to turn smallpox into a more potent bioweapon. "My concern is that this work would break the taboo against genetic engineering of smallpox virus," Tucker says.

Advisers to the WHO committee point out that they are asking for only one genetic modification, which will not change the virulence of the virus. "What we're asking for is limited in scope," says Bernard Moss, head of the Laboratory of Viral Diseases in Bethesda, Maryland. Moss was a special adviser at the WHO meeting. "These modifications will result in the elimination of variola virus earlier," he says.

But Tucker says that if the work goes ahead, the WHO advisory committee will need to strengthen its supervisory capabilities. "The committee really needs a full-time staff person and more resources if it's going to be expanding the smallpox research programme," Tucker says.

Asian biologists seek self-sufficiency in network development

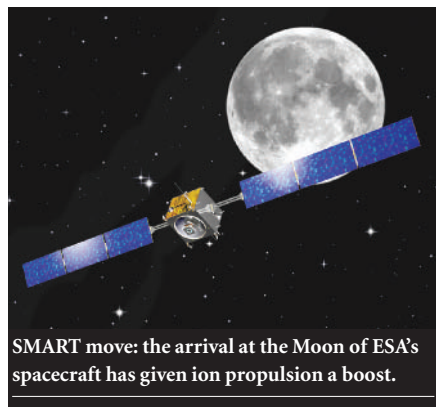
Tokyo More than a dozen prominent developmental biologists from the Asia-Pacific region gathered last week in Kobe, Japan, to launch their own regional scientific network.

The Asia-Pacific Developmental Biology Network will promote student training and the exchange of scientific information, resources and staff from New Zealand to China and as far west as Iran. The network, to be based at the RIKEN Center for Developmental Biology (CDB) in Kobe, plans an inaugural meeting next year at the International Society of Developmental Biologists' congress in Sydney, Australia.

Planners are hopeful that the network will help Asian countries to look to each other for cues, rather than always turning to the United States. "It's been a trend in this region to look to the West first," says CDB director Masatoshi Takeichi, who was appointed as the organization's first chair. "We're trying to encourage people here to get to know what others are doing in their own part of the world as a way of fostering more local collaborations."

Leisurely Moon cruise adds thrust to deep-space flight

Washington For its first trip to the Moon, the European Space Agency (ESA) chose the slow route. What took the Apollo astronauts just three days has been a 13-month marathon for the spacecraft dubbed SMART-1 (Small Missions for Advanced Research in Technology). The craft finally arrived in lunar orbit on 15 November.



SMART move: the arrival at the Moon of ESA's spacecraft has given ion propulsion a boost.

Thankfully this wasn't a race, and researchers say speed isn't meant to be one of SMART's best features. Nor are its science instruments, although its visible and infrared cameras, mineral-mapping spectrometer, and radio and X-ray devices will study the Moon for six months after the craft enters its final lunar orbit in January.

Instead, SMART is showcasing several new

Italians splash out to complain about reforms

Turin Scientists from the University of Turin in Italy flooded onto the streets with squeegies last week to clean car windscreens in a protest against proposed university reforms.

The changes, put forward by Italy's research minister, Letizia Moratti, include the abolition of the academic category of 'researcher', the only position for academics below a professorship.

The reforms were approved by the government in January and will be discussed in parliament next month. This could give them the final stamp of approval.

Italian scientists have spent the summer staging



a range of innovative protests, including a funeral for universities and a lecture series in a pizzeria.

technologies, including an ion propulsion system that spits out a high-speed stream of xenon atoms to give a small but steady thrust. The successful arrival at the Moon has proved that spacecraft could use the same engines to get to Mercury and Mars, says Bernard Foing, who heads the SMART-1 team at ESA's Space Research and Technology Centre in Noordwijk, the Netherlands.

Contest sets tongues wagging for science

London Budding science communicators who fancy a few minutes of fame are invited to audition for FameLab, a competition to be held in Britain next March and April.

Contestants will have three minutes to impress the judges with what the organizers say should be an "entertaining, engaging and informative talk for a non-scientific audience". Twelve entrants will be selected for a final at the Cheltenham Science Festival in June 2005. The winner will be given broadcasting time on Channel 4 television and a UK tour of speaking events.

"Good communication is essential to maintaining confidence in science," says Paul Nurse, president of Rockefeller University in New York and one of the competition's patrons. "If we don't talk about science there may be no science to talk about."

Kids and chemicals study halted for external review

Washington The US Environmental Protection Agency (EPA) has temporarily suspended a study of children's exposure to pesticides and household chemicals following criticism from environmental groups who claim the research is unethical.

The study came under fire in part because it planned to use \$2 million from the American Chemistry Council, a major chemical-industry lobbying group. Environmental watchdogs claimed this presented a conflict of interest, and also

questioned the study's methodology (see *Nature* 432, 6; 2004).

In response, the EPA's acting science adviser, William Farland, announced that the agency would send the study out for an external review by academics from several independent advisory boards. "The EPA is taking this extraordinary step because protecting the health and well-being of children is of paramount importance," Farland wrote in a memo to EPA employees on 8 November. The committee is expected to report back in spring 2005.

Clerics helped to focus on the end of Ramadan

London Muslim clerics in Iran on the lookout for the new Moon last weekend were given official permission to use telescopes in their search by Iran's supreme leader, Ayatollah Ali Khamenei.

Astronomical observations are needed to help determine the end of Ramadan, a month-long daylight fast that starts and ends with a new Moon. Many Islamic followers are happy to base these dates on textbook predictions. But the more traditional approach, in which a religious leader must verify local sightings of the Moon, still persists in many regions, including parts of Iran.

Efforts to spot the first sliver of the Moon can be hampered by dust storms, low cloud or the failing eyesight of ageing clerics. The end of Ramadan — and the start of the three-day Eid al-Fitr holiday — was marked 24 hours apart in some areas of Iran last year, after disagreement over a sighting of the Moon from the holy city of Qom.

Correction

The News story "Beta-blocker goes on trial as asthma therapy" (*Nature* 432, 7; 2004) incorrectly states that beta-blockers reverse high blood pressure by acting on blood vessels. In fact, the drugs act on a number of organs including the kidneys, heart and brain.

Father of fractals

Benoit Mandelbrot is one of the twentieth century's best known mathematicians. So why, in the twilight of an extraordinary academic career, is he still angry with many of his colleagues? Jim Giles investigates.

"I never learnt the alphabet or times tables," says Benoit Mandelbrot of his early education in Poland. Instead, it was visual memories that shaped his early years. He still remembers the geometric patterns that covered the rug on which he took his first steps. And when his uncle began teaching him from home, rote learning was never used. "All I did was play chess and read maps," he says.

His family could not have known it, but they were laying the foundation for a tumultuous career in mathematics — one dominated by remarkable insights and no small amount of feuding. Mandelbrot's skills of visual analysis created new directions in his field: without him, the word 'fractal' would not exist and the usefulness of the bizarre shapes the term encompasses may not have been recognized. But that schooling also produced an academic nomad, long unable to find a home amid the territorial disciplines of maths and physics, and often angry because of it.

Mandelbrot, who celebrates his 80th birthday this month, is now widely revered. Yet a fury born of a sense of injustice still seems to simmer below the surface. Throughout his career, some researchers have questioned the importance of his achievements; Mandelbrot has responded with public explosions of anger and written rebuttals. At an age when others might spend their time accepting accolades or reminiscing with colleagues, Mandelbrot is still working — and angrily defending that work.

From the start, his academic career was unorthodox. After a 1952 PhD in mathematical linguistics, he delved into everything from the theory of financial markets to observations about the true lengths of coastlines. His methods were often unusual. Rather than prove theorems, as most mathematicians do, Mandelbrot used patterns and graphs to make conjectures about physical laws.

Take the paper that eventually put him on the map¹: a 1963 work on the fluctuations of cotton prices that he now describes as his "big bang". At the time, most economists thought that such price changes followed a bell-shaped curve called a gaussian distribution. In effect, they assumed that markets

made frequent small jumps in value and that big fluctuations were rare.

Yet when Mandelbrot looked at charts of cotton prices, he noticed an odd phenomenon. If the label was removed from the time axis, it was impossible to tell whether the charts covered one week or one year; the pattern of peaks and troughs looked the same at each scale. Mandelbrot, then at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York, knew that such 'self-similar' systems follow a different distribution, known as a power law. Crucially, big jumps in value are far more common in these distributions. Mandelbrot showed that the same was true for cotton prices — and in doing so he helped to change the way that stock market firms manage risk.

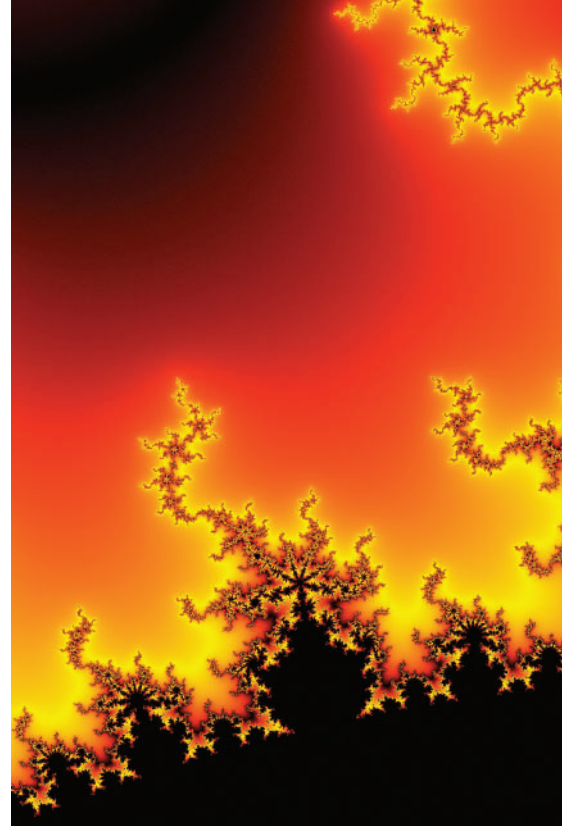
The wanderer

The 1963 paper is now considered a classic, yet its conclusions were disputed at the time. And with few other mathematicians using similar methods, Mandelbrot often had to publish his work in low-impact journals. By his own admission, his limited English and unusual notation also made his work difficult to read. "He almost has his own language," says Kenneth Falconer, a mathematician at the University of St Andrews in Scotland.

So during the 1960s and 1970s, Mandelbrot remained an academic wanderer. Although IBM gave him a base, his diverse interests — from cosmology to geology — deterred university departments from offering him a permanent position. "I was everyone's favourite visiting professor," he says. Stints at Yale and Princeton brought prestige, but not the widespread recognition he wanted.

All that changed in the 1980s. Realizing that most scientists were unaware of work he published in obscure journals, Mandelbrot began to bring together his ideas in a book. When the English version² appeared in 1982, the worlds of maths and physics took notice.

At the heart of his book are fractals — beautifully complex shapes that can be produced from simple mathematical equations. Self-similarity is just one of their remarkable



Shaping up: Benoit Mandelbrot (right) changed the face of maths with his work on fractals, such as the set that bears his name (above).

properties. Take the best-known fractal, since dubbed the Mandelbrot set (see above). Zooming into the sea-horse shapes at the edges of the set reveals more miniature Mandelbrot sets. Jump down a magnification level and the same patterns emerge again.

Largely thanks to Mandelbrot, this property is now known to be widespread in the natural world. Coastlines contain similar patterns when viewed on maps of very different scales. A crack snaking along a metal surface can be fractal, as are the branches of human arteries. Even cauliflowers show some degree of self-similarity.

Repeat performance

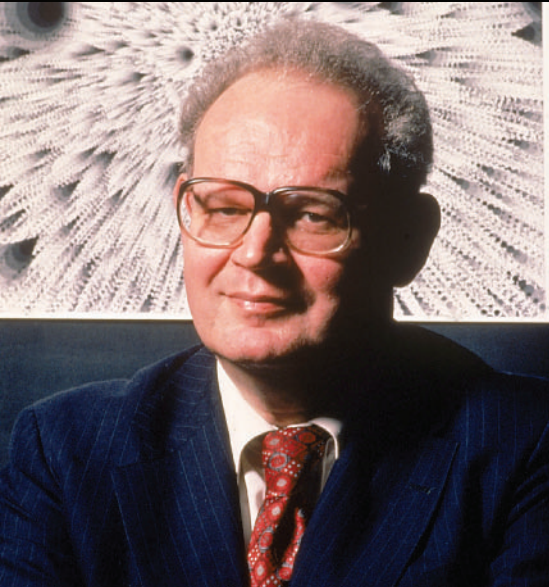
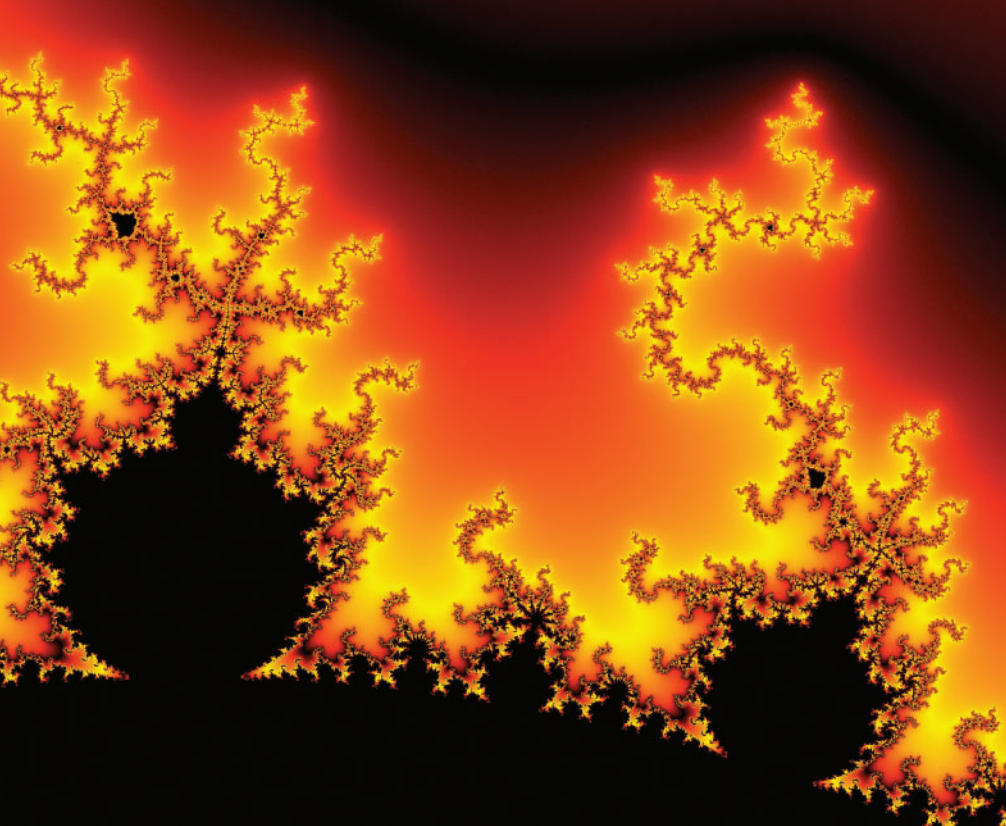
Mandelbrot had wanted recognition, and here it was. Physicists working on topics from cloud formation to metallurgy began using fractals in their work. His book quickly generated hundreds, and eventually thousands, of citations. And stunning plots of fractals, made possible by new computer technology, featured in everything from greetings cards to *Star Trek* movies.

But it is around this time that Mandelbrot developed a reputation for being confrontational. As so often happens in academia, questions of precedence were central. No one denies that Mandelbrot single-handedly put fractals on the scientific map. But in many of the examples he cited in his book, other researchers covered at least part of the ground before him.

Similar studies to Mandelbrot's work on power laws in economics, for example, were written up by Italian economist Vilfredo Pareto at the beginning of the last century³.

"Mandelbrot has changed the questions mathematicians ask."

— Ian Stewart



Lungs are just one example of natural structures that show the self-repeating nature of fractals.

And the famous Mandelbrot set may not have been first plotted by Mandelbrot — Robert Brooks and Peter Matelski, mathematicians then at the University of Maryland and the State University of New York at Stony Brook, respectively, did so at about the same time⁴.

When this latter case was pointed out in a book review in 1989, Mandelbrot penned a rebuttal claiming that Brooks and Matelski had plotted a “crude version” of the set and had given “no thought” to its special nature⁵. And that is far from the only time he has made such comments about other researchers’ contributions. “He can get up in the middle of other peoples’ lectures and claim to have done the same work years ago,” says one mathematician who has worked on fractals. “He can be really quite aggressive.”

Gene Stanley, a theoretical physicist at Boston University, Massachusetts, has expe-

rienced some of that aggression. An argument between Stanley and Mandelbrot brought a 1996 conference on fractals, held at New England College in Henniker, New Hampshire, to a temporary standstill. Researchers who were present say that the row, which centred on a dispute over who would chair the next conference, degenerated into an all-out shouting match. One delegate was prompted to cry out, tongue-in-cheek: “Why can’t we all just get along?”

Most academics can recall similar feuds in their own fields, albeit probably less fiercely fought ones. But Mandelbrot, who has been at Yale since 1999, has been accused of more unusual behaviour in relation to his *Selecta* books — a collection of reprints of Mandelbrot’s original papers. When the books are compared to those originals, differences emerge.

In his “big bang” paper on stock-market statistics, for example, many references to Pareto have been removed from the reprint⁶. The “Paretian hypothesis”, which refers to Mandelbrot’s new way of thinking about the

statistics of markets, becomes the “L-stable hypothesis”, after the French mathematician Paul Lévy, who worked on the same problem. Another theorem, the Pareto–Doebelin–Gnedenko conditions, is in some places renamed the Doebelin–Gnedenko conditions. And one reprint, contributed by Eugene Fama, an economist at the University of Chicago in Illinois, has had its 1963 title changed from “Mandelbrot and the stable paretian hypothesis” to “Mandelbrot on price variation”⁶.

Time for heroes

So has Mandelbrot attempted to write Pareto out of his papers? One researcher, who asked not to be named, has made just this accusation to Springer, the Berlin-based publishers of the *Selecta* books. But Mandelbrot denies that this is the case, saying that Pareto, whose achievements he acknowledges at length in his latest book⁷, is one of his heroes. He adds that the changes were made so that the terminology was consistent. Fama was not aware of the changes to his paper, but backed Mandelbrot’s explanation when *Nature* made him aware of the editing.

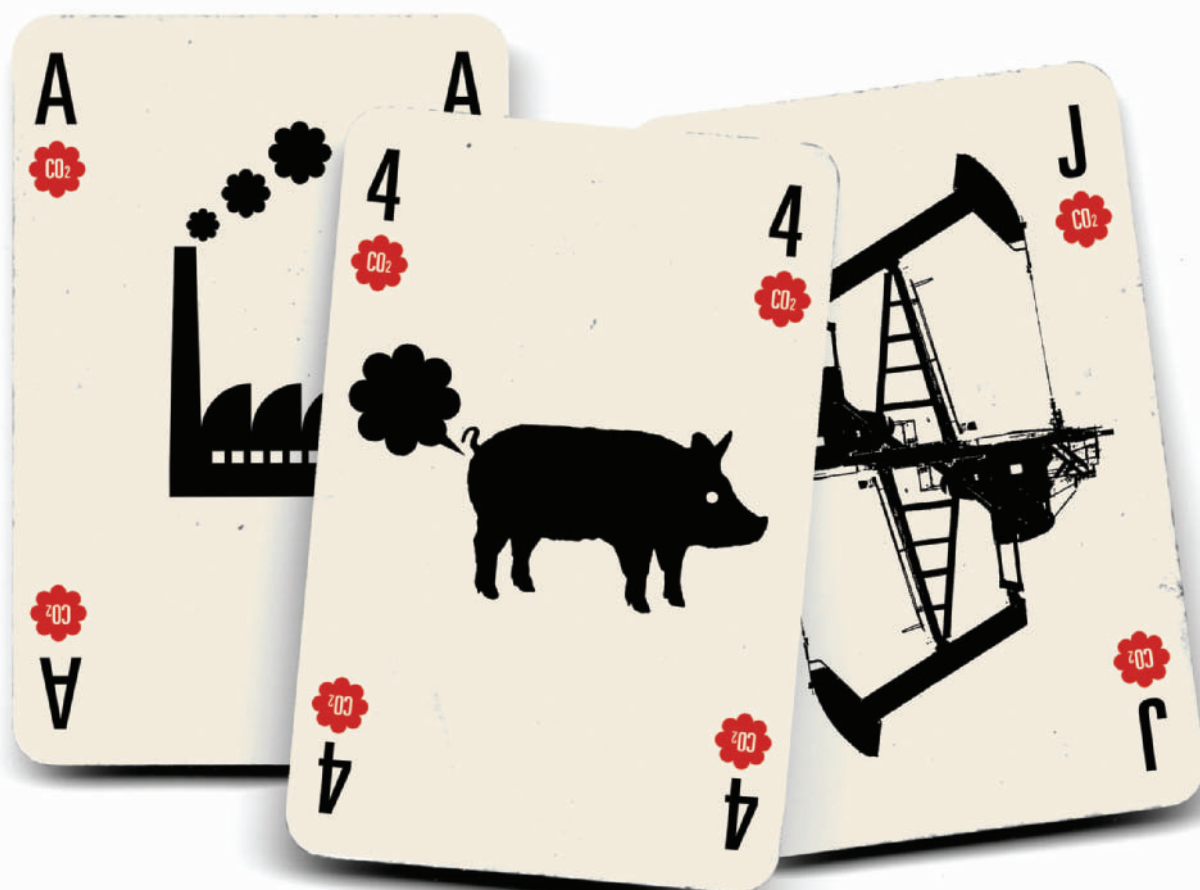
Despite these controversies, Mandelbrot’s colleagues are generally more interested in what he has achieved than how he may have behaved. Some feel fractals are over-hyped, but the majority say that their introduction has changed the way physicists think about natural phenomena — for which they owe Mandelbrot a considerable debt.

Even researchers who have been the subject of Mandelbrot’s attacks praise his contributions to maths. And several physicists told *Nature* that they altered the focus of their research after hearing him speak. Ian Stewart, a mathematician at the University of Warwick, UK, sums it up: “He has changed the questions we ask.”

Hearing such praise earlier in his career might have mellowed Mandelbrot. But ask him today whether he has received the recognition he deserves, and he delivers a typically combative reply. He says that many new books on finance omit to mention his work, and colleagues still pursue grudges against him. Is this normal behaviour for a distinguished mathematician in the twilight of his career? Perhaps not. But, as Falconer points out, Mandelbrot is not a normal mathematician. ■

Jim Giles is a reporter for *Nature*, based in London.

1. Mandelbrot, B. B. *J. Bus.* **36**, 394–419 (1963).
2. Mandelbrot, B. B. *The Fractal Geometry of Nature* (W. H. Freeman, New York, 1982).
3. Pareto, V. *Manuale di Economia Politica* (Società Editrice Libreria, Milan, 1906).
4. Brooks, R. & Matelski, J. P. in *Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference* (eds Kra, I. & Maskit, B.) 65–71 (Princeton Univ. Press, Princeton, 1981).
5. Mandelbrot, B. B. *Math. Intelligencer* **11**, 17–19 (1989).
6. Mandelbrot, B. B. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk* (Springer, New York, 1997).
7. Mandelbrot, B. B. & Hudson, R. L. *The (Mis)behaviour of Markets: A Fractal View of Risk, Ruin and Reward* 152–159 (Profile Books, London, 2004).



The CARBON game

Companies are already swapping money for the right to emit more pollution, and cashing in on projects designed to suck up greenhouse gases. As this market booms, will it actually help to cut down on emissions? Michael Hopkin reports.

This summer, a group of power companies in Japan and Canada developed an unusual interest in pig manure. The porcine waste was at the heart of a landmark multimillion-dollar deal between Chile's largest pork producer and the power companies, allowing the latter to emit more pollution. The pig farm promised to recycle its animals' emissions of methane — a potent greenhouse gas — by covering the manure, capturing the gas, and burning it as sustainable energy. In return, the power companies bought the right to emit more carbon dioxide from their stations, half a world away from the oblivious pigs.

The deal, signed in August, is one of the largest such exchanges, but it is by no means the first: trading in greenhouse-gas emissions has been going on since the mid 1990s. That may seem strange in a world that has not yet entered into the full grip of the Kyoto Protocol — the international agreement that aims to reduce greenhouse-gas levels, in part by allowing companies to buy and sell 'equivalents' of carbon dioxide emissions. But, as the ink is drying on that deal thanks to Russia's recent ratification of the treaty (see *Nature* 431, 1030; 2004), carbon markets are already doing brisk business.

Carbon trading has been slowly growing since its inception (see Graph, opposite), and is now set to explode under the stewardship of a handful of dedicated brokers. By the end of this year, the total volume of CO₂ traded is

expected to be double that of 2003. But 2005 will be the year in which the trade truly comes of age, when in January the European Union (EU) launches its Emissions Trading Scheme (ETS), involving some 12,700 industrial organizations spread across all 25 EU member states. By 2007, the European market is expected to be worth €10 billion (US\$13 billion) per year, says Henrik Hasselknippe, an analyst at consultancy firm Point Carbon in Oslo, Norway.

Trading places

Emissions limits in the ETS are established by National Allocation Plans — proposals submitted by each EU member state that are now being individually approved by the European Commission. Once they get the green light, governments farm out their allowances to industrial installations — such as power companies, mineral miners, and cement and paper manufacturers — giving each an allotment of 'emissions credits' that they can trade internationally. The governments involved keep track of the emissions, based mainly on the known inputs to these installations, and update their figures with information on registered trades. For every tonne of CO₂ emitted above the limit, companies face a fine of €40, rising to €100 from 2008 onwards.

In tandem with schemes designed to increase compliance with Kyoto — such as the ETS and forthcoming systems in Canada

and Norway — companies also trade in emissions on their own, in part to showcase a certain amount of civil responsibility to their stock-holders. In the absence of official, nationally allocated 'carbon emission credits', they trade in projects instead — such as the pig-farm deal. Some trading even goes on in countries that have not ratified the Kyoto deal, such as the United States and Australia. And independent day-traders, who will use the market to make a quick buck rather than to manage emissions, are expected to arrive on the scene soon.

Carbon trading is seen by most as an economic necessity — one that might also benefit the environment. Turning emissions into a commodity gives companies a financial incentive to clean up their act by more than is legally required: a small, voluntary market set up in the 1990s in the United States seems to have proved that emissions fall when reductions are made valuable. But some researchers and politicians are sceptical that a global scheme will likewise reduce emissions, rather than simply shifting pollution around. The negotiations that set up national carbon allowances may have been over-generous, they say, the price of carbon credits may be too cheap, and the future reliability of plans to generate credits may be suspect.

Gas exchange

The idea of a carbon market was written into the Kyoto Protocol thanks mainly to pressure from the United States, which nevertheless pulled out of the treaty in 2001. The idea was that many countries would agree to long-term emission reductions only if companies were allowed to buy the right to more emissions when needed. Such a trading scheme would spread the burden of emissions costs to those most able to support them. This has been shown to work in other contexts, primarily the US scheme for trading sulphur dioxide emissions, which is widely regarded as a triumph of market forces enforcing environmental rules.

The US scheme, known as the Acid Rain Program, was a drive to cut emissions from coal-burning power stations in response to a 1990 tightening of the US Clean Air Act. The programme worked on a 'cap and trade' basis for SO₂: more than 260 of the most polluting power stations were given a maximum emissions allowance; anyone destined to exceed their limit had to buy extra credit from those who were ahead of the curve. By 2000, when the market was opened up to all US coal-fired power stations, SO₂ emissions nationwide had fallen to about 11 million tonnes per year, down from almost 16 million tonnes in the late 1980s¹.

The ETS will attempt to pull off a similar trick for Europe's CO₂ emissions. But with

25 countries all striving to cut emissions by about 1% each year, it is a much more formidable proposition than the US scheme, which involved trading within a single industrial sector in a single country. "No one has ever done a trading scheme on this scale before," says Neil Strachan, an economist and climate researcher at the Pew Center on Global Climate Change in Arlington, Virginia.

Observers fervently hope that the ETS is a success — not just as an emissions-reducing measure but as a demonstration of Europe's traditional leadership in tackling climate change. If it works, it will set an example to the rest of the Kyoto countries, which in 2008 must begin fulfilling their own commitments.

Most trading in the ETS will fall into one of two categories: straight buying and selling of credit between members, and investment in the Clean Development Mechanism (CDM), a United Nations scheme to encourage the growth of environmentally friendly industry in the developing world.

The CDM has already thrown up some innovative ideas — the Chilean pigs again being an example. A handful of other strategies have also passed the UN's stringent verification process, which is designed to ensure that they deliver the carbon credit they promise. These include destroying the greenhouse gas HFC-23, a by-product in the manufacture of some chemical refrigerants, or converting the fumes from landfill sites into energy. But the most widely cited potential CDM method — planting forests of fast-growing trees such as eucalyptus to soak up CO₂ — has yet to be approved. Regulators have yet to be persuaded that schemes of this nature are safe bets.

Critics say the outcome of CDM projects that involve sequestration of CO₂, rather than limiting emissions, can never be guaranteed.

"What happens when you plant a forest, then there's a change of government and they chop it down and build a holiday resort? Do you then destroy the power plant that got built because of the project?"

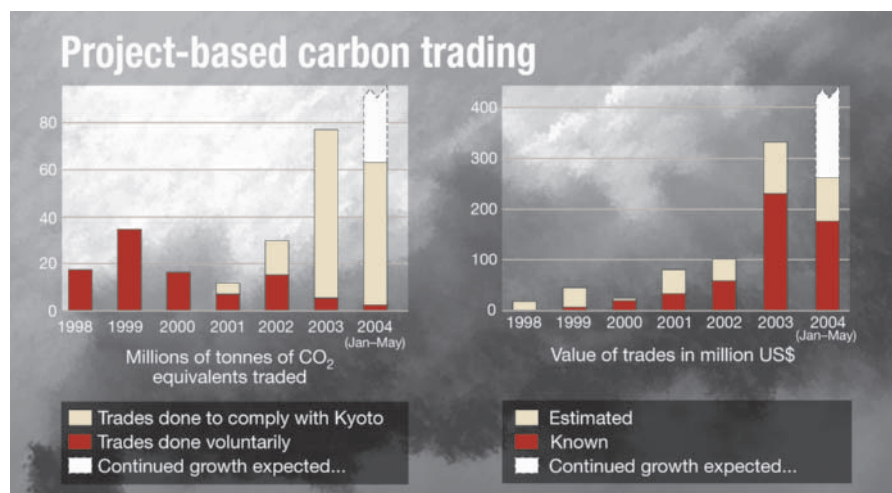
— Michael Dorsey

At the moment there are no accurate ways to predict the change in CO₂ levels from such projects, or to verify that these changes actually take place. Even if there were, it would still leave other concerns. "What happens when you plant a eucalyptus forest, then there's a change of government and they chop it down and build a holiday resort?" asks Michael Dorsey, an environmental scientist at Dartmouth College in Hanover, New Hampshire. "Do you then destroy the power plant in America that got built because of the project?" Without carbon deficits as well as credits — and without an effective way to penalize those who don't play by the rules — Dorsey argues it is nearly impossible to ensure that emissions will go down instead of up.

In credit

Others say that the entire CDM mechanism simply allows Western industry to keep developing while dropping stopgap solutions on the doorstep of the developing world. Although the projects may bring some industry and money to these nations, it isn't everyone who is benefiting. Currently, more than two-thirds of emissions reductions are supplied by just five countries: India, Brazil, Chile, Romania and Indonesia². Most African countries have yet to broker a single deal. And once Russia enters the scene — which it is expected to do soon — it should be able to earn billions from its emission quota, taking a large chunk of the market.

Russia is likely to have huge amounts of credit on its hands because Kyoto targets are based on emissions in the treaty's baseline year of 1990. Back then, smokestack industries of the old communist regime were still belching out masses of CO₂. Since the collapse of the Soviet Union in 1991, Russia has shut these plants down, giving itself a surplus of allowable emissions that it can now sell on. Critics of carbon trading point out that if Russia sold all of its credits the consequences for emissions would be effectively the same



Growth industry: the trade in 'carbon credits' that buy companies the right to emit more pollution has increased steadily — by the end of this year, trade is expected to have doubled over 2003.

SOURCE: REF. 2

as if someone cranked these old power-houses back up to speed. In the short term that would be bad news for emissions, although the world would still be better off than it was 14 years ago.

Perhaps of greater concern are the emissions expected from other areas of industry that are exempt from the Kyoto Protocol and excluded from the carbon markets set up to help regulate it. Emissions from cars and planes, for example, are not included in national allocations. "The trading system covers only about 50% of total EU emissions — other sectors, such as transport, are still growing," says Hasselknippe.

Fair price

Despite these criticisms, the prevailing sentiment among marketeers and environmentalists is that emissions trading is the only pragmatic way forward, particularly considering how well it worked for SO₂. After all, that market likewise faced complaints in its early years. "At the time we started SO₂ trading we were called smog-traders and garbage-peddlers — Greenpeace even picketed us," recalls says Richard Sandor, chief executive of the Chicago Climate Exchange, which was involved with the project. "But it has saved lives." He says that the \$1.2-billion cost of the scheme to the US economy has bought a \$27-billion reduction in the healthcare burden of lung disease through cuts in acid rain.

Sandor is now setting his sights firmly on a single worldwide carbon market, and has launched the European Climate Exchange, which will broker deals between members of the ETS. The move comes after the success of a voluntary carbon-trading scheme run by the Chicago exchange, also overseen by Sandor. This US market currently involves some 75 members drawn from the power, manufacturing, forestry and agriculture sectors, including names such as Ford, IBM, DuPont and Rolls-Royce. Together, they produce more emissions than the whole of Britain. Members signed a contract to reduce emissions by 1% a year, but Sandor claims that since trading began in December 2003, members' overall emissions have gone down by 9% — proof, say advocates, that the market works in favour of the environment.

The incentives for these US players are very different from those that relate to Kyoto. Some polluters may want to take action to avoid lawsuits such as those currently dogging the tobacco industry, says Sandor. Others may want to invest profitably in an environmentally friendly industry. Still others may want to steal a march on any



Bad hand: critics claim the carbon game can only succeed if it includes emissions from transport.

legislation that may appear in the future — or rules that already exist. In Oregon, for example, power plants have an obligation to reduce emissions to 17% below that of the most efficient combined-cycle plants, or pay \$0.85 per tonne of excess emissions².

Encouragingly, regulations in the voluntary Chicago scheme — which were determined by consulting businesses about what levels of reductions they would like to aim for — are surprisingly similar to regulations in the ETS, which were handed down from on high by the European government: both arrived at targets of 1% reduction a year. If industry self-assessment matches government enforcement, many argue, a workable worldwide market is a realistic possibility, regardless of different countries' motivations for playing the carbon game.

Once such a world market is established, it should, ideally, be self-regulating. The harder it is to comply with nationally enforced regulations, the higher the price of credits and the harder some companies will work to supply them.

At the moment, the price of a tonne of

CO₂ credit in Europe is €4–5 for a CDM investment and €8–9 for a more straightforward trade of credits with another market member. Market observers say that these prices are quite low — slightly lower, for straight trades, than they were as recently as this February, thanks to relatively lenient caps given to many European companies by some national allocation plans. But the price could do anything once the market gets going, says Reena Qureshi of London-based co2e.com, which brokered the pig-farming deal. "It will respond to interest rates, or some analyst saying it's going to move, or political crises like the Iraq war," Qureshi says. Even the weather could sway the market — another heatwave like the scorching summer of 2003 would send power use skyrocketing, meaning that power companies could need to buy more emissions credit.

Share and share alike

Once it is officially launched, the European market will also be open to — and is expected to catch the attention of — independent day-traders hoping to make a quick buck. Suddenly the state of world emissions will become of intense interest to those who sit at their computers swapping stocks and shares. If nothing else, the market could then prove a positive exercise in increasing awareness of global warming.

Regardless of the price that credits reach, some observers say that it will always be too cheap compared with the true cost of pollution. "There's no mechanism for bringing prices into line with the social cost," says Melissa Carrington, a London-based analyst at consultancy firm PricewaterhouseCoopers. The fines — €40 per tonne for ETS members — will place a ceiling on market prices that many say is too low. One British government study puts the overall social cost of climate change to society — in terms of flooding, extreme weather and health issues — at about €100 per tonne of carbon emitted³.

Still, advocates point out, the market offers an economically viable strategy to get businesses on track towards emissions reduction: without carbon trading, the Kyoto Protocol would never have passed. As it is, the market puts companies in the right frame of mind, knowing that reductions equal cash. This in turn encourages power companies to manage their emissions and pig farms to recycle their manure — and surely that's a good thing.

Michael Hopkin is a reporter for news@nature.com based in London.

1. Ellerman, A. D., Joskow, P. L. & Harrison, D. *Emissions Trading in the US: Experience, Lessons and Considerations for Greenhouse Gases* (Pew Center on Global Climate Change, Arlington, VA, 2003).
2. Lecocq, F. *State and Trends of the Carbon Market 2004* (Carbon Finance at the World Bank, Washington DC, 2004).
3. Clarkson, R. & Deyes, K. *Estimating the Social Cost of Carbon Emissions* Government Economic Service Working Paper 140 (HM Treasury, London, 2002).

Let's be sensible about public participation

We must face the fact that science — like art — is not a democratic activity.

Sir — Your Editorial “Going public” (*Nature* 431, 883; 2004), like the think-tank Demos, supports the fashionable demand by a group of sociologists for more democratic science, including more ‘upstream’ engagement of the public and its involvement in setting research priorities. Demos goes further and supports a ‘needs test’ for licensing new products or services by companies. It also argues that we, the public, should know who owns and controls new technologies, and who benefits, before they are developed.

If the Demos policy had been followed in the past, we would have neither electricity nor the laser, to name only two examples, because no practical uses were foreseen for either. As your Editorial admits, public-engagement exercises in the United States have led patient lobby groups to press the National Institutes of Health

for less basic research and more drug development. Because of public demand, large sums are spent on developing drugs with Viagra-like properties rather than on medicines for people in developing countries, and a widespread public consultation exercise in Oregon has found strong opposition to spending limited public funds on AIDS or mental health.

In practice, greater involvement of ‘the public’ in the ‘upstream’ development stage of science means involvement of special-interest groups. When the UK Agriculture and Environment Biotechnology Commission was set up, the ‘public’ representatives were the chair of Greenpeace, the chair of the Soil Association, the executive director of GeneWatch and the programme adviser to the Green Alliance. No wonder the ‘GM Nation’ exercise in public consultation was a fiasco.

Of course democratically elected governments must decide how public funds for science are allocated. Of course sensible consultation helps development of policy: the debate on stem-cell research in the United Kingdom was a good example. Of course more openness and transparency are to be encouraged where possible. But let us not display unthinking subservience to the principle of participation. In Britain, involvement by victims of rail accidents in deciding policy on railway safety has led to the investment of billions of pounds to save some five lives a year. Meanwhile, twice that number die on British roads every day. The fact is that science, like art, is not a democratic activity. You do not decide by referendum whether the Earth goes round the Sun.

Dick Taverne

Chair of Sense About Science, House of Lords, London SW1A 0PW, UK

Public participation: let the people pick projects

Sir — Your editorial “Going Public” (*Nature* 431, 883; 2004) makes a persuasive case for upstream public engagement in science funding. No doubt setting up committees of non-scientists to advise the existing funding bodies is a step in the right direction. But there is also a more radical possibility, namely to set aside a small proportion of the public science budget, say 1%, for research proposed by lay people.

What questions would be of public interest? Why not ask? Organizations such as charities, schools, local authorities, trades unions, environmental groups and gardening associations could be invited to make suggestions. Within each organization, the very possibility of proposing research could trigger far-ranging discussions and would lead to a sense of involvement in many sections of the population.

To avoid the 1% fund being taken over by the science establishment, it would need to be administered by a board largely composed of non-scientists, as in many research charities. Funding would be restricted to areas not already covered by the other 99% of the public science budget.

This system could be treated as an experiment and tried out for, say, five years. If it had no useful effects, it could be discontinued. If it led to productive research, greater public trust in science

and increased interest among students, the percentage allocated to this fund could be increased.

Rupert Sheldrake

20 Willow Road, London NW3 1TJ, UK

Bible study led Newton to scientific discoveries

Sir — I feel that your News story “Newton’s religious screeds get online airing” (*Nature* 430, 819; 2004) rather misses the point. To our modern minds, Isaac Newton’s religious ideas may indeed seem “unorthodox” or “radical”, but they did not look like this to his contemporaries. Like another father of modern science, Francis Bacon (see “A modern kind of magic”, *Nature* 418, 821; 2002), Newton strongly believed that he lived in an era that had been predicted by the Book of Daniel of the Old Testament, a time when knowledge was expected to grow beyond recognition.

Throughout his life, Newton tested biblical truth against the physical truths of experimental and theoretical science. He never observed a contradiction. The order that he found in nature through experiment and calculation — later to be called the mechanistic worldview — was for him God’s work, and proof of God’s work in history, which he extracted from the Bible. Astronomical calculations helped him to synchronize biblical events described in the Old and New Testaments

with what he knew about ancient, medieval and modern history.

To English Protestants during the seventeenth century, when the country was consumed by apocalyptic zeal, the Book of Daniel and The Apocalypse, or The Revelation of St John, were history — revealed truth — even though they were written in visionary and symbolic language. The task was to turn these visions and symbols into modern language.

A whole host of scientific writers — including the illustrious Cambridge polymath Joseph Mede — took to the task of interpretation. It was on the shoulders of these giants that Newton was standing when he wrote his main religious work *Observations upon the Prophecies of Daniel and the Apocalypse of St John*, which was intended as an update of world history based on the five-kingdoms scheme in the Book of Daniel.

It is probably one of the deepest ironies in the history of science that Newton’s brilliant work did not serve the purpose that he intended. Rather than proving the Bible right, it led to the birth of science as we know it — that is, experimental natural science.

We now know that the Book of Daniel’s five-kingdoms scheme is a myth and The Revelation of St John is a wonderful fairy tale. But it is the Bible, nevertheless, that stands, in a very literal sense, at the origin of modern science.

Erwin Heberle-Bors

Vienna Biocenter, Dr Bohrgasse 9, A-1030 Vienna, Austria

Time for 'enlightened moderation'

A call for Islamic nations to renew and reaffirm their commitment to science.

Atta-ur-Rahman and Anwar Nasim

The time has come for a renaissance in the Muslim world, for a new strategy of 'enlightened moderation'. In the wake of 11 September 2001, Islamic countries face myriad challenges and the gap of misunderstanding between the West and the Islamic world is widening. The way forward for Muslim countries is, in their own interest, to focus on internal reforms and socio-economic modernization, to shun extremism and to promote moderation.

The global security situation has given Islam a false image, that of a religion of intolerance, activism and terrorism. Islam is unfairly linked with fundamentalism, fundamentalism with extremism, and extremism with terrorism.

Muslims can argue all they like that this loose thinking is unfounded, but we are having little impact in today's battle of ideas. It doesn't help that some Muslim nations are probably among the poorest, the least educated and the least powerful on the planet. We must get out of this rut if we do not want to be marginalized and to condemn future generations.

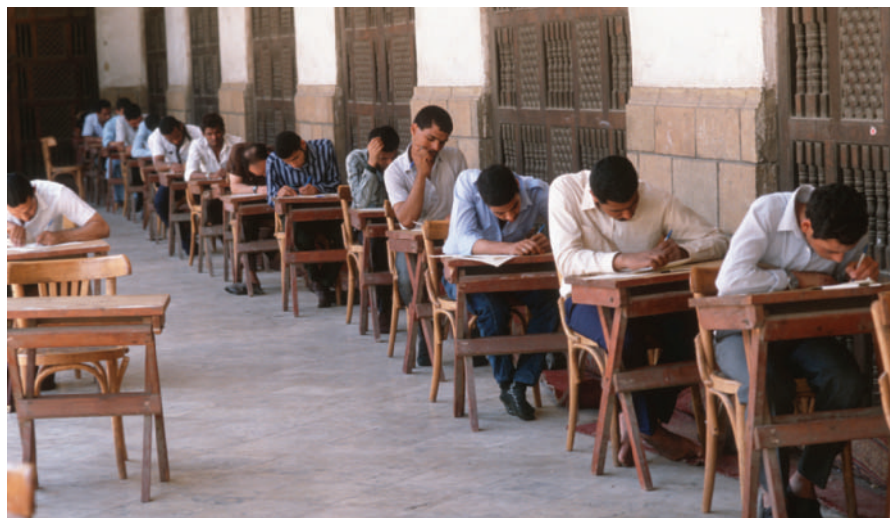
The Organization of the Islamic Conference (OIC) is a group of 57 geographically scattered countries with predominantly Muslim populations. Stretching from Indonesia to Morocco and from Uganda to Kazakhstan, they are home to 1.3 billion people, but their economies are generally among the world's poorest, and illiteracy levels are among the highest. Six of the eight poorest countries on the planet are OIC members.

This is a time for critical thinking and soul-searching among Muslims — in particular for leaders of Islamic nations, who can play a key role in bringing about "enlightened moderation", as envisioned by Pakistan's president, Pervez Musharraf. We in Pakistan believe that science and technology are crucial to a knowledge-based renaissance.

There is no shortage of ideas and proposals. History will not forgive those who are at the helm of affairs today, but who fail to respond with enlightened policies and actions to shape the destiny of over a billion people. Bold initiatives in science must form part of any response, and the deep resurgence of science in Pakistan over the past five years, as described below, shows what can be achieved.

An exemplary society

Science and Islam share a glorious past. In its heyday, Islam was the standard bearer of a society of law and order, justice, tolerance



Science in higher-level education in many Islamic countries is underfunded, leading to a lack of research.



Al Azhar University in Cairo is the oldest in the world. It covers both science and religious studies.

and exemplary values. The Koran encourages the pursuit of science, and the Islamic world was a cradle of science from the eighth to the fifteenth centuries.

The Muslim world of today has strayed far from those values. We have fallen behind in socio-economic development and in the generation of ideas. During our decline, we have shut ourselves off and refused to absorb knowledge from others. And our spending on science is dire. We can regret this deplorable situation, but we also need to face up to it.

The West can help to usher in an era of enlightened moderation by contributing to the planning and funding of centres of excellence in the Islamic world, where mutual understanding and tolerance could flourish. But in this article we will focus on what the Muslim world needs to do. We need to ask ourselves some tough questions: as Muslims, what are our ideas? Where are we going?

Will confrontation and political activism bring us back to our glorious past? No. We

must take an enlightened path dedicated to developing our human resources, and tackling the problems of poverty, education, health and social justice. We must abandon confrontation in favour of moderation, conciliation and individual freedom. It is time for renaissance of the Ummah (the global Muslim community). This is how we will eliminate the perception of Islam in conflict with modernity and democracy.

Science in Islamic states

One might argue that science is universal, and that it is impossible to speak separately of science in Islamic states, such as Pakistan, or Asian science. But by any comparison of international science indicators, Muslim states emerge as a well-defined cluster with many common characteristics and needs.

Political leaders in many Islamic nations largely fail to appreciate the importance of scientific research to their countries' development. Public spending is often skewed towards the military, educational standards are low and public interest in science is undeveloped.

For many Muslim countries, all the socio-economic warning lights are flashing, be it in terms of literacy, poverty, or the quantity and quality of scientific workforces and their output. The situation, to put it bluntly, is dismal.

The Islamic world's average science spending is at an order of magnitude below global averages¹. In contrast, spending on defence averages from 4% to 7% of the GNP. In many countries, the population of scientists is meagre and legal frameworks for innovation are largely non-existent. Only two scientists from Islamic states have won Nobel Prizes, Abdus Salam, a Pakistani

(Physics, 1979) and Ahmed Zewail, an Egyptian (Chemistry, 1999). Both carried out their research outside Islamic countries. Today's Muslim societies have generated few scientists of international repute.

Take science education. Many of the Arab OIC member countries, including Malaysia, have fairly good undergraduate education systems, but are weak at the postgraduate level. Of the world's top 500 universities, only two are in OIC member states (both in Turkey). The OIC's 1.3 billion inhabitants are served by less than 600 universities, most of low standard.

Making way

But progress is being made. Turkey, for example, ranked 46th in the world in terms of output of scientific papers seven years ago; it jumped to 22nd place in 2002, according to Philadelphia-based science information specialist, Thomson ISI. The international isolation of Iran over the past two decades has generated a strong sense of self-reliance and autonomy from the West — and a resulting increase in investment in science and technology.

Since 1999, Pakistan has increased science spending 60-fold, and funding of higher education 12-fold. Funds have been released for 1,500 PhD students to be trained annually at home, and a further 300 abroad. As a result, the annual PhD output should increase from about 200 this year to 1,200–1,500 by 2009.

A nationwide digital library of 20,000 journals has been launched, providing free access for all educational and research institutions. There is also a new scheme to attract 1,500 top researchers back from overseas to work in Pakistan during the next five years. In addition, the cabinet has launched plans to reinforce research and make science a priority of long-term policy at cabinet level.

Turkey leads OIC states in terms of annual output of research papers with 6,393 in 2001, then Egypt at 2,498, followed by Iran² — which has tripled its output from 501 in 1996 to 1,830 in 2002. During the period 2001 to 2003, the sharpest increase has come from Pakistan, with a 40% increase from 636 to 890. This is a result of a system introduced in 2002 that provides researchers with an opportunity to more than quadruple their earnings if they increase the numbers of their papers published in peer-reviewed journals.

Global effort needed

These pockets of improvement are encouraging. But the fact remains that OIC countries are home to three-quarters of the world's fuel reserves and a quarter of its other natural resources. How do we reconcile this richness in resources with our lack of socio-economic development? Our backwardness in science and technology and higher education is part of the answer.



Standing up for science: Abdus Salam (above, left) and Ahmed Zewail (above), are the only Nobel laureates to come from Islamic countries. Ekmeleddin Ihsanoglu (left) has vowed to improve relations in international research. Pakistani president, Pervez Musharraf has revitalized spending on science.



and bioinformatics. Due to the paucity of funds, the bank has agreed to earmark only US\$450,000 at present.

How can the West help to strengthen science in Islamic countries? The Inter-Academy Panel on International Issues — an alliance of 90 scientific academies — organized a workshop in 2003 in Trieste, Italy. This brought together research ministers and heads of scientific societies to promote the creation of more independent scientific academies in Muslim countries, with the goal of boosting both research and independent scientific advice available to governments³.

A follow-up meeting in Islamabad in March this year resulted in the creation of the Network of Academies of Science in countries of the OIC.

To the future

Science is a truly global activity, and although the OIC and COMSTECH have their role to play, encouraging bilateral and international cooperation is the key to progressing towards enlightened moderation in the Islamic world.

This is just a beginning in the task of redressing centuries of neglect by our political leaders. For the policy-makers, the writing is on the wall — there is a need to develop a knowledge economy, face the challenges of the new world order and spend at least 1% of GNP on strengthening science and technology. The OIC member states must respond to all these challenges.

Encouragingly, Ekmeleddin Ihsanoglu — appointed last June to head the OIC for four years — has pledged to improve cooperation between researchers from Muslim countries and others worldwide. Ihsanoglu is also president of the International Union for the History and Philosophy of Science. He needs to muster the necessary political and financial support to revitalize science and technology for socio-economic development in OIC member countries.

As we continue to dream, struggle and search for a happier and more prosperous future, we share this new awareness among Islamic states with the global scientific community. We urgently seek your much-needed cooperation and interaction. In the place of the clash of civilizations, our collective wisdom and efforts can help heal wounds and guarantee a safer and better world for those who will follow us. ■

Atta-ur-Rahman is the Federal Minister & Chairman of the Higher Education Commission, and Coordinator-General, COMSTECH.

Anwar Nasim is Science Adviser to COMSTECH. COMSTECH Secretariat, 3 Constitution Avenue, G-5/2, Islamabad 44000, Pakistan.

1. Human Development Report (UN Development Programme, New York, 2004); <http://hdr.undp.org/reports/global/2004>.
2. Science Watch 14, (2003); www.sciencenewswatch.com/nov-dec2003/sw_nov-dec2003_page1.htm.
3. Butler, D. *Nature* 422, 101–102 (2003).

Immune to the facts

A flawed paper on autism compromised MMR vaccination and public health.

MMR Science & Fiction: Exploring A Vaccine Crisis

by Richard Horton
Granta: 2004. 220 pp. £7.99

MMR and Autism: What Parents Need to Know

by Michael Fitzpatrick
Routledge: 2004. 232 pp. £45, \$75(hbk);
 £14.99, \$23.95 (pbk)

Michael B. A. Oldstone

The measles virus is remarkably contagious, infecting more than 95% of susceptible humans exposed to it. About 1 in 1,000 people infected requires hospitalization and may become permanently disabled, and 1 in 300,000 infected with measles develops a progressive neurological disorder, subacute sclerosing panencephalitis, which invariably causes death.

But along with smallpox, yellow fever and poliomyelitis, measles is remarkably well controlled by vaccination. Before routine measles vaccination in the United States, some 4,000,000 people a year contracted the disease; widespread vaccination has brought this number down to less than 40. Other countries in which 95% of the population are vaccinated — the proportion believed to provide 'herd immunity', which blocks the transmission of measles — have experienced similar declines in cases. However, coverage in some regions of Africa is less than 50% of the population, and there are still some 40,000,000 infections each year, causing more than 600,000 deaths.

These figures highlight the amazing achievements of measles vaccines and the need for mandatory vaccination to maintain this record of success. Nonetheless, the anti-vaccination sentiment, which began after Edward Jenner first used cowpox vaccine to prevent smallpox, continues today. The reasons are complex and multifactorial but often arise from basic libertarianism or individualism, along with people's desire to stop the government and public-health bodies having so much control over them and their children. Others consider vaccines harmful or refuse to face even the minor risk of side-effects. Many people are misled and others misinformed.

This was the backdrop to a recent attack in Britain on the combined measles, mumps and rubella (MMR) vaccine by some parents with autistic children, by news reporters and by some in the British government. As a result, MMR vaccination coverage sank to 85% or below in Britain, soon to be followed by outbreaks of measles and, sometimes, devastating after-effects. So the publication



Afraid of needles? Fewer children receive the MMR vaccine since scare stories linked it to autism.

of two books reviewing autism and the search for an environmental cause, especially the MMR vaccine, are particularly timely.

MMR Science & Fiction by Richard Horton, editor of *The Lancet*, and *MMR and Autism* by Michael Fitzpatrick, a general practitioner of medicine and parent of an autistic child, both make the case that the MMR vaccine is safe and is not associated with autism. Both books provide the background to help readers understand the events in Britain and the players involved, as well as the evidence that supports their conclusion that the vaccine is safe.

Autism, first recognized in the 1940s, results in children having an inability to relate to themselves or to people and situations. Its cause is unknown. However, in 1998, *The Lancet*, with Horton as editor, published a paper by the gastroenterologist Andrew Wakefield and colleagues at the Royal Free Hospital in London that linked autism and the MMR vaccine. According to the paper, of 12 children who had both autism and chronic enterocolitis (bowel disease), eight had been given the vaccine and one had a measles infection before the onset of autism. No virological evidence was provided of the measles infection, nor was there any stringent epidemiologic evidence to link it, or other events, to the autism, and there was no control group of children. Fuelled by concern and self-interest, some of the authors, journalists and parents of autistic children went to the newspapers, citing the article as

proof of a connection between MMR and autism. The publicity sparked a public movement against vaccination that quickly grew out of control.

Of the two books, Fitzpatrick's provides the more thorough and detailed account of the events as they unfolded. It includes a critical evaluation of the flawed paper in *The Lancet* and discusses autism in greater depth. This book is a good read.

Horton's book contains similar information but focuses primarily on the events, complications and failures that followed the publication in *The Lancet*. As an antidote, he suggests that an ombudsman group of scientists and laymen should evaluate such controversial findings to prevent the spread of misinformation. He also describes the profiteers: journalists and publicists, medical practitioners who received fees from lawyers representing the parents of the autistic children, and physicians and scientists who were compensated by pharmaceutical houses, biotech companies and government officials. Horton attributes the public mistrust to these sources. These financial conflicts of interest, along with questions over informed consent for the children, resulted in several, but not all, of the authors retracting the paper six years after its publication. Horton does not deny his own responsibility, stating that if he knew in 1998 what he knows now, *The Lancet* "would not have published the part of the paper that related to MMR".

But the crux of the issue is missing from

Horton's book. Without firm evidence of measles infection or of the measles virus in the MMR vaccine being involved in disease, why was the claim of an association between autism, the measles virus and the MMR vaccine placed in the results section of the paper? In Horton's defence, the same issue of *The Lancet* published a commentary pointing out the defects in Wakefield's paper, but the question remains: why did *The Lancet* open Pandora's box?

Good science demands objectivity based on experimental or clinical evidence that is reliable and reproducible. Subjective opinion is not proof and has no place in the peer-reviewed literature. Indeed, as quoted in Horton's book, Britain's chief medical officer, Liam Donaldson, stated: "If the paper had never been published, then we wouldn't have had the controversy, we wouldn't have had the seed of doubt sown in parents' minds which has caused a completely false loss of confidence in a vaccine that has saved millions of children's lives around the world." Donaldson is correct.

What's more, the continued focus of Wakefield, journalists and parents of autistic children on autism's link with measles, coupled with the complacency of a public that has been shielded from the horrors of uncontrolled infections, has supported a continued antivaccination movement. This has had a regrettable influence on the public and the political establishment concerning measles and autism specifically, and vaccination in general. ■

Michael B. A. Oldstone is head of the Division of Virology, Department of Neuropharmacology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

Across the border

Alien Species and Evolution: The Evolutionary Ecology of Exotic Plants, Animals, Microbes, and Interacting Native Species

by George W. Cox

Island Press: 2004. 377 pp. \$75 (hbk), \$40 (pbk)

David M. Lodge

While some governments are preoccupied with preventing border crossings by terrorists, thousands of alien species (those from other regions or continents) continue to be allowed free entry into most countries of the world. Some of these alien species are certain to cause great harm to the environment, native species, national economies and human health, as other species have done in the past. In this era of supposedly great attention to border security, how do the aliens keep on getting through?

Commercial markets in live food, pets,



Alien invader: the Surinam toad was introduced to Australia from its native South America.

horticulture and aquaculture intentionally import a wide range of alien species into many countries every year, with little government supervision and often no analysis of the attendant environmental, health and financial risks. Thousands of other species hitch-hike on legitimate cargo or the ships, planes and other vehicles that carry them.

Most of these aliens will do little or no harm, but some will cause irreversible damage. Recent examples include the North American grey squirrel in Europe; the Asian longhorned beetle in North America; the Northern Pacific seastar and the Surinam toad in Australia; the European red deer in South America; the South American water hyacinth in Africa; and the Australian brown tree snake on Pacific islands. Changing patterns of trade mean that increasing numbers of alien species come from previously isolated regions.

The environmental and economic damage wrought by alien species includes the extinction of native species, and large alterations in ecosystem characteristics, such as nutrient fluxes and fire frequency. Charles Elton anticipated many of these environmental effects in his 1958 book *The Ecology of Invasions by Animals and Plants*. Indeed, it often seems that invasion biology (a recent addition to the list of biological specialisms) is little more than Elton redux. However, in *Alien Species and Evolution*, George Cox extends traditional concerns about alien species beyond the ecological theatre, and puts the evolutionary play on centre stage. His main concern is genetic change, both in alien species, which are subject to founder effects and new selection pressures, and in native species, as they experience new selection pressures imposed by the aliens.

This extremely readable book is aimed primarily at students and researchers. Cox

provides comprehensive coverage of alien species in different taxonomic groups and in different habitats: terrestrial, freshwater and marine. Replete with examples and abundantly referenced, the book provides an excellent evolutionary synthesis. Cox occasionally makes extended forays beyond alien species, but only to illustrate the broader context in which adaptation and counter-adaptation occur. The book is therefore also a good introduction to the broader intellectual landscape of evolution and global environmental change.

The coverage of hybridization between alien and native species may be particularly useful, as many readers might not have encountered it before. The interaction between hybridization and polyploidy has already produced a number of new terrestrial plant species from ancestral species, for instance when European salsifies (*Tragopogon*) were introduced into North America. Hybridization and introgression have also been common in freshwater fishes, crustaceans and molluscs, as human interventions have brought closely related species together. This often results in the loss of native species as an evolutionary and ecological entity, as well as a chance to study evolution in action.

Cox provides a guide to other research topics where an understanding of evolution is essential, including the development of invasion resistance by native communities. He considers evolution during lag times of invasion, which arise because different life-history characteristics are often required for dispersal and persistence in a particular environment. Cox also provides an introduction to coevolution between alien and native species, and to the effect of invasion on geographic speciation — although he makes it clear that this result is relatively minor compared with extinctions caused by invasions.

Little explicit attention is given to policy responses to species introductions, but Cox illustrates the need for greater consideration of evolutionary processes in risk analyses for alien species. In particular, he gives several examples of hybridization between wild species and related crops that have been genetically engineered for resistance to herbicides or insect pests. Transgenes have already flowed (or almost certainly will if they haven't already) from cultivated sorghum into Johnson grass, from oilseed rape into field mustard, from sunflower crops into wild sunflowers, and from wheat into jointed goatgrass. Wild species do not seem to suffer any reduction in fitness from incorporating some transgenes, contrary to the claims of many proponents of genetic engineering, so there are likely to be detrimental effects on native insects. Cox convincingly makes the case that evolution is central to any understanding of invasions, and that the analysis of risk is incomplete

without a consideration of evolution. ■
David M. Lodge is at the National Center for Ecological Analysis and Synthesis, Santa Barbara, California 93101, USA. He is on sabbatical from the University of Notre Dame, Notre Dame, Indiana 46556, USA.

Exhibition

And all was light

The Newtonian Moment: Science and the Making of Modern Culture

Curated by Mordechai Feingold
At The New York Public Library until
5 February 2005.

Alan Packer

Isaac Newton is rarely out of the news these days. Last year there was the publication of *Isaac Newton*, James Gleick's elegant distillation of Newton's character and science. Last month came the final instalment of *The Baroque Cycle*, Neal Stephenson's rollicking trilogy of novels, in which Newton plays a pivotal role. Now he gets the scholarly treatment in *The Newtonian Moment: Science and the Making of Modern Culture*, an exhibition at The New York Public Library.

Curated by Mordechai Feingold of the California Institute of Technology, the exhibit presents maps, prints, books and models from the library's collection, and manuscripts from the Cambridge University library. Newton's death mask, once owned by Thomas Jefferson, is also on display. The narrative focuses on Newton as "innovator and icon of the Enlightenment", and traces the reception of his ideas in their historical context.

Who would not be moved by the sight of one of Newton's early notebooks, remarkably well preserved, or an early edition of the *Principia Mathematica* (1687) with handwritten notes for proposed revisions? Bibliophiles will enjoy the different editions of the *Principia*, and *Opticks* (1704), translated into a variety of languages. Some of these are adorned with a dramatic allegorical frontispiece celebrating Newton's mastery of celestial and terrestrial mechanics and his revelatory insights into light and colour.

But if the exhibition emphasizes the apotheosis of Newton, it makes clear that his was not an easy road to glory. For every Voltaire — whose *Elements of Newton's Philosophy* was one of the most successful popularizations of newtonian thought — there was a *Celestino Cominale* (*Anti-Newtonianism pars prima*), or an Etienne Simon de Gamaches, who held fast to his compatriot Descartes' theory of vortices in the face of Newton's theory of universal gravitation. And there

was a long-running dispute between Newton and Leibniz over which of them deserved the credit for discovering the calculus.

Feingold also presents the texts and drawings of important figures whose work and world-views were affected by Newton: William Blake and his illustrations of reason and imagination; Alexander Pope ("God said Let Newton be! and All was Light"); as well as Immanuel Kant, Johann Wolfgang von Goethe and Denis Diderot. All of these were forced to confront the revolution in mathematical rigour that Newton had launched.

Newton's passion for mathematics also fuelled his idiosyncrasies. Accompanying his diagrams is a floor-plan of Solomon's Temple, whose dimensions he painstakingly derived from biblical descriptions. Newton kept this obsession, among others, carefully concealed during his lifetime. ■

Alan Packer is senior editor at Nature Genetics.



Page proofs? A first edition of *Principia Mathematica* with Newton's handwritten notes on the left page.

A walk on the wild side

Out of this World: Colliding Universes, Branes, Strings, and Other Wild Ideas of Modern Physics

By Stephen Webb
Copernicus: 2004. 308 pp. €29.95, £17.50, \$27.50

Frank Close

"This book is about some really wild ideas," is how *Out of This World* begins, so readers cannot say they have not been warned. I lost count of the number of times that 'wild' was juxtaposed with 'ideas', although other adjectives, including 'outrageous', 'phenomenal' and 'outlandish' are also used to describe the "amazing recent theories" on hidden dimensions, branes and modern particle physics.

The book is about high-energy physics; if readers have enough energy of their own, they might pick up some of the sense of excitement in the current research, but they shouldn't expect suddenly to see the light. I remember seminars where the speaker attempts to review the field in the first three

minutes to set the stage for the presentation; those who know the field don't need it, and those who don't aren't likely to learn much. Too often I was left with this feeling as I ploughed through this book: we are given brief glimpses of the standard model, grand unification theories and supersymmetry, superstrings and more.

Stephen Webb tries to describe the panoply of particle physics without using equations, which is a noble aim. He must have read widely to achieve this, although on occasion his sources can be rather visible; an allusion to the strength of the weak force being like Cleopatra falling off her barge in 50 BC but not yet hitting the water rang a bell, for example. The book falls short of existing books, such as Brian Greene's *The Elegant Universe* (W. W. Norton, 1999), which covers much of the material with greater assurance. There is too much in Webb's book that raises doubts about the reliability of the material and suggests that *Out of This World* is out of its depth.

Here is a sample, by no means exhaustive. The book states that a particle with "zero rest (*sic*) mass" always travels at the speed of light. Poorly drawn or incorrectly labelled diagrams violate electric charge, illegally convert quarks into leptons, or show the strength of the weak force, an SU(2) structure, strengthening at energies above 100 GeV, whereas it actually starts to become feeble again.

There is the wild idea that "on average, two quarks in a free neutron come close enough to exchange a W boson about once every ten minutes". If the author is equating the ten-minute half-life of the neutron with the chance that the W boson can initiate the "weak" beta decay, this is misleading. The typical lifetime of a particle resisting beta decay is of the order of nanoseconds or less: the longer lifetime of the neutron is mainly due to the fact that the combined masses of proton and electron, into which the neutron decays, are so close to that of the free neutron that there is almost no 'phase space' available.

To popularize using metaphor requires the author to have a deep understanding of the material. To do this over some 300 pages without using equations is a challenge of the highest order, so it is perhaps not surprising that Webb does not always succeed. Readers might have found it easier to persevere if he had shown less ambition, following the maxim that 'less is more'. The book does provide a sense of the development of ideas, and how the frontiers of current mathematical particle physics are developing, but the descriptions are patchy and the "wild ideas" rather overstated. Without a sense of irony, the dust-jacket puff reads: "Then, in a series of increasingly astonishing chapters...". Astonishing indeed! ■

Frank Close is at the Rudolf Peierls Centre for Theoretical Physics, 1 Keble Road, Oxford University, Oxford OX1 3NP, UK.

Going against the flow

Ion transport: the division between active transporters and passive channels is beginning to blur

Louis J. DeFelice

To function properly, living cells must maintain ions and small molecules at concentrations that are far different from those in their local environments. The electrochemical gradients formed in this way can be used by cells as a source of energy to drive other processes, for example membrane transport or the generation of electrical signals in specialized cells — such as those in nerves and muscle. Internal sodium ions are kept at roughly one-tenth of their external concentration, and the reverse is true for potassium ions. These ion gradients are maintained by a transmembrane protein (the sodium–potassium pump, an enzymic transporter), which actively pumps out sodium ions and pumps in potassium ions. Then, as an electrical signal (action potential) propagates along a nerve cell, sodium and potassium channels open, allowing the rapid flow of sodium ions into and potassium ions out of the cell. From many such examples, the concept emerges that transporters generate ionic gradients and channels dissipate them: transporters and channels seem as different as oil and water.

But this distinction depends on what we classify as ‘transporter’. For enzymatic transporters, the difference seems clear — they use chemical energy to generate gradients. But for another class of transporters, known as co-transporters, the distinction is more difficult. Co-transporters can also create gradients, but they do so by using the energy stored in ionic gradients established by pumps — and are thus secondarily active. For example, co-transporters would use the sodium gradient set up by the sodium pump to power the transport of another substrate, such as a neurotransmitter, against its own gradient. However, researchers tend to model co-transporters as if they were enzymes, even though they have authentic channel properties. What are the origins of this riddle and how can it be resolved?

Channels and transporters did not originate as separate entities. Alan Hodgkin and colleagues unravelled the ionic basis of action potentials in the 1940s. But to explain sodium and potassium permeability they initially postulated transporters, not channels — ions were carried across the membrane rather than moving through a pore. Soon afterwards, prominently in the work of Hodgkin and Richard Keynes, the notion of electrodiffusion through narrow pores (channels) emerged. Concurrently,



researchers struggled with electrodiffusion as a means to concentrate metabolites ‘against the gradient.’ Wilfred F. Widdas, in particular, introduced the harbinger of co-transport, in which the gradient of one species drives the transport of another. The adoption of enzymatic theory and carrier kinetics for co-transporters soon followed, and two camps evolved: electrodiffusion theory governed channels, and enzyme theory governed transporters. Different methodologies contributed to this separation, as channel biophysicists relied mainly on electrical measurements, whereas transport physiologists preferred radiolabelled uptake experiments. Indeed, early attempts at an electrical description of active transport were disappointing because of the comparatively low signal.

As a result, the study of transporters (including co-transporters) and the study of channels grew apart. But in the past decade the cloning of co-transporters, combined with the measurement of tiny currents through individual channel and transporter proteins by high-resolution electrophysiology (patch clamp) has begun to bridge this historical gap. Almost every co-transporter studied in this way exhibits ion channel properties. Glutamate and dopamine co-transporters harbour chloride-selective channels. GABA, serotonin and norepinephrine co-transporters contain sodium and lithium channels. Recently, a presumed chloride channel has been shown to be a co-transporter, but a simple mutation returns it to pure chloride selectivity. In another case, an iron co-transporter naturally mutates into a calcium channel. We may expect many other such examples as the application of

ion channel techniques are applied to transporters. Even the sodium–potassium pump has been shown to have ion-channel properties under special conditions.

The challenge now is to reconcile the signature property of a co-transporter — the ability to accumulate substrates ‘against the gradient’ — with its nature as an ion channel. At least two possibilities arise. Co-transporters might obey enzyme theory when they are active, but occasionally slip into a passive channel mode. Alternatively, co-transporters might obey channel theory and rely on flux coupling for their secondary activity.

In flux coupling, an ionic current drives the secondary transport of a substrate. This may occur because a channel is too narrow to allow the ion and substrate to pass one another and so the powerful flow of the ion down its gradient carries the substrate along against its own gradient (single file diffusion). Electrical and osmotic forces can also drive flux coupling — for example, an ionic current can drive water against its own osmotic gradient and vice versa. On a cautionary note, flux coupling in the same direction seems simple and obvious, but a much knottier problem is that of counter flow, in which one species moves in the opposite direction to the other and yet coupling is still positive.

We seem closer than ever to understanding similarities between these seemingly disparate membrane proteins, due to the advent of structures for several ion-selective channels. These include a chloride–hydrogen co-transporter (first presumed to be a chloride channel), lactose and glutamate co-transporters and an extremely high-resolution solution for an ammonia channel (once thought to be a co-transporter). Releasing co-transporters from the grip of enzyme theory would give us a fresh start on mechanisms that would include all transport properties, and lead to a possible reunion of co-transporters and channels. ■

Louis J. DeFelice is in the Department of Pharmacology and the Center for Molecular Neuroscience at Vanderbilt University Medical Center in Nashville, Tennessee 37232, USA.

FURTHER READING

DeFelice, L. J. *Trends Neurosci.* **27**, 352–359 (2004).
Hodgkin, A. L. *Chance and Design* (Cambridge Univ. Press, 1992).

Lauger P. *Electrogenic Ion Pumps* (Sinauer Associates, 1991).

Stein, W. D. *Channels, Carriers, and Pumps* (Academic Press, New York, 1990).

At the root of brain cancer

Michael F. Clarke

A small subpopulation of cells, 'brain-cancer stem cells', has been identified in humans. They have the exclusive ability to drive tumour formation, and could prove an effective target for therapies.

Brain cancers are among the most devastating tumours in humans and are often rapidly fatal despite aggressive treatments. These tumours typically contain varied populations of cells that differ in the specific proteins — or markers — displayed on the cell surface.

On page 396 of this issue, Singh and colleagues¹ describe how they have isolated a minority population of human brain-cancer cells based on the expression of a cell-surface marker called CD133. They report that, when injected into the brains of mice, this subpopulation of CD133⁺ cells could by itself drive tumour growth and dissemination. As few as one hundred of the CD133⁺ cells formed tumours that could be serially transmitted from mouse to mouse, whereas tens of thousands of cancer cells lacking CD133 failed to do so. When tumours that arose from the injected CD133⁺ cells were examined, the cellular heterogeneity and architecture closely resembled that of the patients' tumours from which the cells had originally been taken. These findings add brain tumours to the list of cancers, including blood^{2–4} and epithelial cancers⁵, in which a cancer stem-cell population has been found.

Insights into how tissues are maintained provide hints as to why only a minority of the cancer cells drives tumour formation. Most cancers arise in tissues, such as the bone marrow, gut and breast, that are composed of a cellular hierarchy in which a small population of stem cells gives rise to progenitor cells that regenerate mature tissue cells. It was once thought that neuron formation in the brain was essentially complete by birth. However, recent findings demonstrate that the brain, like other organs in which cancers arise, contains a stem-cell population that can give rise to differentiated cells, including mature neurons.

In the normal brain, neuronal stem cells as well as early progenitor cells, but not their fully mature progeny, express the CD133 marker (Fig. 1a)⁶. In the brain tumours examined, Singh *et al.*¹ found distinct subpopulations of cells that expressed either CD133 or various markers of mature brain cells. Thus, the cellular architecture of the brain tumours may be a caricature of that of the normal brain, with brain-cancer stem cells, probably derived from normal CD133⁺ stem or progenitor cells, giving rise

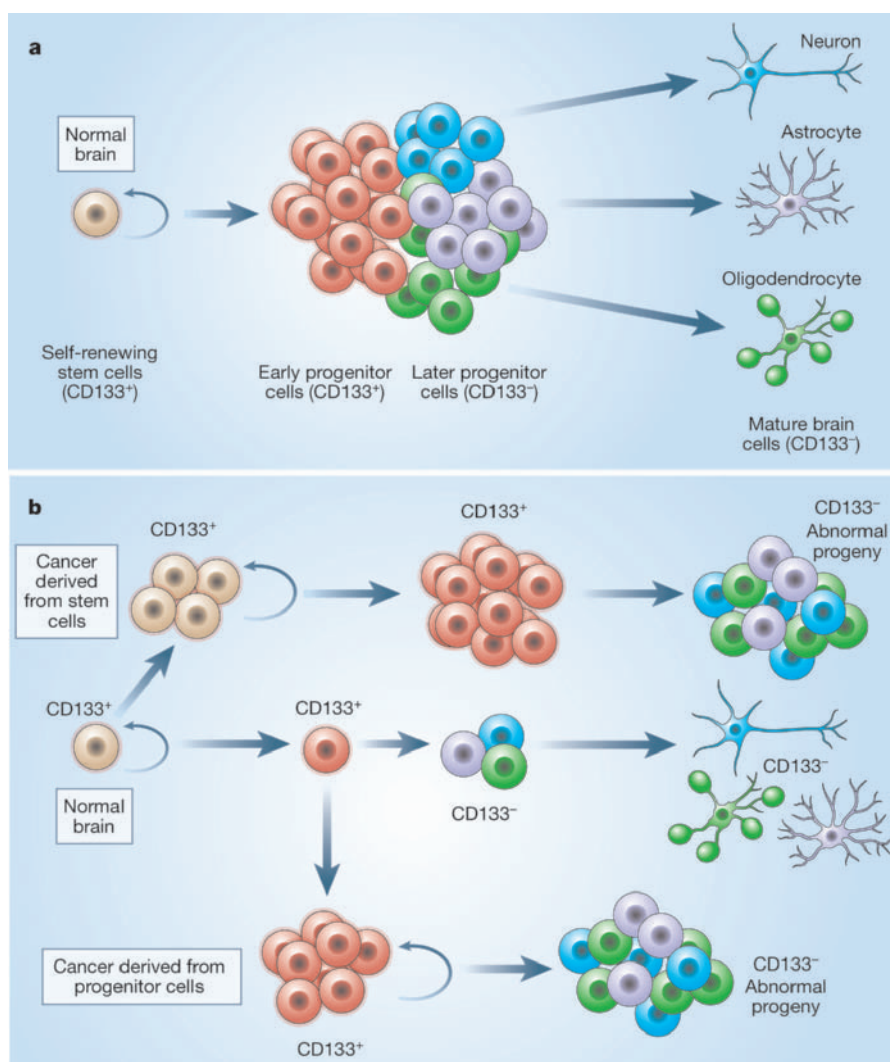


Figure 1 Brain-cell hierarchy. a, In the normal brain, stem cells, which express the CD133 marker and so are designated CD133⁺, can generate new stem cells by the process of self-renewal. They can also produce early progenitor cells (CD133⁺), and later progenitor cells (CD133⁻) that give rise to the mature forms of brain cells (neurons, astrocytes and oligodendrocytes, all CD133⁻). Unlike the stem cells, progenitor cells have limited ability to replicate. b, Singh *et al.*¹ have identified cancer stem cells. Such cells could arise from CD133⁺ brain stem cells, when loss of normal constraints results in expansion of the abnormal stem-cell pool, or from early CD133⁺ progenitor cells as a result of mutations that make these cells self-renewing. In either case, the cancer stem cells also generate abnormally differentiated CD133⁻ progeny that cannot self-renew and thus cannot form new tumours.

to aberrantly differentiated progeny (Fig. 1b).

Stem cells have two unique properties that make it likely that they are involved in cancer development. First, they are often the only long-lived cells in a tissue that have the ability to replicate. Typically, multiple mutations, occurring over many years, are necessary before a cell becomes cancerous. So

the implication here is that cancer-inducing mutations accumulate in the long-lived, normal stem cells.

Second, through a process called self-renewal, stem cells generate new stem cells with similar proliferation and differentiation capacities to their parental cell. By contrast, with each round of replication,

progenitor cells become progressively more differentiated and are eventually destined to stop proliferating. Predictably, self-renewal is an essential property of some cancer cells, and at least some genes that regulate normal stem-cell self-renewal also do so in cancer cells^{7–9}. This suggests that cancers arise either from normal stem cells or from progenitor cells in which self-renewal pathways have become activated (Fig. 1b).

Using serial transplantation, Singh *et al.*¹ demonstrated that at least some of the CD133⁺ brain-cancer cells can renew themselves. By contrast, the ability of CD133[−] cancer cells to proliferate was limited. Soon after injection of the CD133[−] cells into the brain, they stopped growing, leaving behind a small cluster of quiescent tumour cells.

This outcome is reminiscent of the observation that many women with small clusters of metastatic breast-cancer cells in their bone marrow, who did not receive any treatment, survived for years without progression of their cancer¹⁰. One explanation for this tumour dormancy is that the microscopic clusters of cancer cells did not contain cancer stem cells and so, like the CD133[−] brain-cancer cells, were unable to grow further. Taken together with the observation that circulating cancer cells in the blood are an indicator of prognosis in breast-cancer patients¹¹, this suggests that the use of markers to reveal cancer stem cells could help in making decisions about treatments.

The identification of cancer stem cells is a significant step in the fight against these

dreaded diseases: because self-renewal is essential if tumours are to grow, agents that target such cells may be effective treatments. A possible complication is that the mechanisms known to regulate cancer-stem-cell self-renewal also regulate the process in normal stem cells. Unlike normal stem cells¹², however, the expansion of cancer stem cells is not tightly regulated, implying that there are significant differences between the normal and the cancerous self-renewal pathways. This gives hope that the isolation of cancer stem cells, coupled with our knowledge of the mutations causing cancer, will result in ways to eliminate cancer cells while sparing normal tissues. ■

Michael F. Clarke is in the Departments of Internal Medicine, and Cell and Developmental Biology, University of Michigan Medical School, CSCG Room 4410, 1500 E. Medical School Drive, Ann Arbor, Michigan 48109-0936, USA.
e-mail: mclarke@umich.edu

1. Singh, S. K. *et al.* *Nature* **432**, 396–401 (2004).
2. Falkow, P. J. *Birth Defects Orig. Art. Ser.* **12**, 123–132 (1976).
3. Lapidot, T. *et al.* *Nature* **367**, 645–648 (1994).
4. Matsui, W. *et al.* *Blood* **103**, 2332–2336 (2004).
5. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
6. Uchida, N. *et al.* *Proc. Natl Acad. Sci. USA* **97**, 14720–14725 (2000).
7. Lessard, J. & Sauvageau, G. *Nature* **423**, 255–260 (2003).
8. Park, I. K. *et al.* *Nature* **423**, 302–305 (2003).
9. van Lohuizen, M., Frasch, M., Wientjens, E. & Berns, A. *Nature* **353**, 353–355 (1991).
10. DiStefano, A., Tashima, C. K., Yap, H. Y. & Hortobagyi, G. N. *Cancer* **44**, 196–198 (1979).
11. Cristofanilli, M. *et al.* *N. Engl. J. Med.* **351**, 781–791 (2004).
12. Morrison, S. J. *et al.* *J. Immunol.* **168**, 635–642 (2002).

the activation barriers and transition-state vibrational frequencies for all the relevant elementary surface reactions; second, to couple these through statistical mechanical methods involving transition-state theory and kinetic Monte Carlo simulations of the reaction process. Vibrational frequencies for all the intermediates and transition states, also calculated quantum mechanically, are used to incorporate entropy considerations into the calculated rate constants. Monte Carlo simulations, which are in widespread use in science, allow a highly complicated system to be sampled efficiently in a number of random configurations, the upshot being a description of the system as a whole. Kinetic Monte Carlo takes this a step further, allowing efficient sampling of the tremendous range of different timescales necessary to describe all the different elementary steps, and thus simulate the kinetics of the whole system.

In this way, Reuter *et al.* were able to calculate net catalytic reaction rates on solid surfaces under conditions similar to those used in industrial processes. Other workers have used a related approach to investigate different types of phenomena², but this application to catalytic reaction rates is a particularly demanding test. Reuter *et al.* have achieved impressive accuracy in calculating the steady-state rates of the catalytic oxidation of carbon monoxide (a reaction performed by the catalytic converters of automobiles) over a model ruthenium oxide (RuO₂) catalyst. Their rates agree almost perfectly with excellent experimental rate measurements, performed by a different group at the same institute³, over a wide range of reaction conditions. The beauty of Reuter and colleagues' method is that it allows one to identify which elementary steps control the reaction rates and how their rates are affected by reaction conditions and, in principle, by surface structure. This is just what is needed to guide the development of better catalysts.

There is, unfortunately, one major limitation associated with the accuracy of quantum-mechanical calculations when applied to chemisorbed species on surfaces. The quantum method used by Reuter *et al.* is a version of density functional theory (DFT) that achieves nearly state-of-the-art accuracy in predicting the energies of such systems. Nevertheless, the authors admit that this can still be off by as much as 30 kJ per mol in estimating activation barriers. Given this drawback, it is surprising that they were able to achieve such impressive accuracy in their calculated carbon monoxide oxidation rates over RuO₂. They attribute this agreement with experimental rates to an effect they imply is generic to catalytic reactions — the combined action of many elementary steps that simultaneously affect the rate.

Although I agree that surface-catalysed reactions generally have many elementary

Chemistry

Towards tomorrow's catalysts

Charles T. Campbell

The ability to predict and modify the rate-determining steps in chemical reactions would be a boon in designing better catalysts. Technical innovations in computer simulations bring that goal closer.

Chemical conversions driven by catalysts are essential to modern society. But we must do better: sustaining industrial and economic growth, while protecting the environment, will necessitate developing more efficient processes. In particular, there is a pressing need for new solid catalysts for mixed-phase, or heterogeneous, reactions. Such reactions tend to involve large volumes, and they are the basis of, for example, more efficient ways of reducing plant and automotive emissions, and of producing cleaner fuels such as hydrogen or methanol.

A way forward lies in predicting how the details of a catalyst's surface structure control key kinetic parameters in the reaction mechanism. One such parameter is the activation barrier, which, if known for the

rate-controlling elementary steps, allows the relevant rates to be calculated. These in turn enable accurate predictions of both the rate of production of the desired products and the branching ratios to undesired products, knowledge of both of which is essential to ensure a successful outcome in terms of energy efficiency and environmental impact. Writing in *Physical Review Letters*, Reuter *et al.*¹ herald a powerful new approach that should ultimately help make prediction possible.

Reuter and his colleagues show just how close science has come to using *ab initio* theoretical methods to calculate net catalytic reaction rates on complex solid surfaces. They have applied an elegant method that they call “*ab initio* statistical mechanics”, which involves two stages — first, using first-principles quantum mechanics to calculate

steps, I doubt that this explanation for the accuracy of Reuter and colleagues' method can be extended to other catalytic reactions. It is well known that there are usually fewer than three rate-controlling steps in such reactions, in spite of their mechanistic complexity. Indeed, under many conditions there is a single rate-determining step⁴⁻⁶. The degree of rate control is usually greater than 0.5 for the most important step in such reactions, and it is unity when there is just one rate-determining step. A degree of rate control of 0.5 means that any (differential) error in the rate constant of that step would lead to an error in the net catalytic reaction's rate by that same factor, scaled by 0.5 (ref. 6). An error of 30 kJ per mol in the activation energy for such a step could thus cause the net mechanism's rate at 500 K to be in error by more than a factor of 500!

It would be interesting to apply the degree of rate control⁶ to the authors' kinetic model so as to estimate which step in this reaction's mechanism has the largest degree of rate control. If its degree of rate control is greater than 0.3, it must mean that the authors were rather lucky in getting the activation barrier for this key step as accurately as they did with DFT, or in having its error cancel with other errors.

Another complication in tackling such complex systems is the difficulty of guessing which of the elementary steps are relevant, so that they can be included in the kinetic Monte Carlo simulation. Sometimes, processes on surfaces occur by rather unexpected elementary steps. Fortunately, this problem has been solved by Henkelman and Jonsson^{7,8}. Unlike the situation in traditional kinetic Monte Carlo, in their method the atoms are not assumed to sit on lattice sites, and a list of all possible transitions need not be specified beforehand. Rather, their method elegantly finds the relevant transitions on the fly during the simulation.

Reuter and colleagues' paper¹ is important: it indicates that the day is not so far distant when computational methods can achieve the necessary chemical accuracy to guide catalyst development. Already, new *ab initio* approaches⁹⁻¹¹ are on the horizon that promise improvements in energy accuracy relative to current DFT methods, with acceptable costs in increased computer time.

Meanwhile, experiments will remain essential: in the foreseeable future, the best microkinetic models for solid-catalysed reactions will continue to be those that also incorporate experimentally measured activation energies for elementary steps. But the theoretical approach of Reuter *et al.* will be an essential complement to overall strategies intended to reveal structure–function relationships in heterogeneous catalysis. ■

Charles T. Campbell is in the Department of Chemistry, University of Washington, Seattle, Washington 98195-1700, USA.
e-mail: campbell@chem.washington.edu

1. Reuter, K., Frenkel, D. & Scheffler, M. *Phys. Rev. Lett.* **93**, 116105 (2004).
2. Ovesson, S., Bogicevic, A. & Lundqvist, B. I. *Phys. Rev. Lett.* **83**, 2608–2611 (1999).
3. Wang, J., Fan, C. Y., Jacobi, K. & Ertl, G. *J. Phys. Chem. B* **106**, 3422–3427 (2002).
4. Boudart, M. *Kinetics in Chemical Processes* (Prentice-Hall, Englewood Cliffs, NJ, 1968).
5. Boudart, M. & Tamaru, K. *Catal. Lett.* **9**, 15 (1991).
6. Campbell, C. T. *J. Catal.* **204**, 520–524 (2001).
7. Henkelman, G. & Jonsson, H. *J. Chem. Phys.* **115**, 9657–9666 (2001).
8. Henkelman, G. & Jonsson, H. *Phys. Rev. Lett.* **90**, 116101 (2003).
9. Kudin, K. N., Scuseria, G. E. & Martin, R. L. *Phys. Rev. Lett.* **89**, 266402 (2002).
10. Staroverov, V. N., Scuseria, G. E., Tao, J. & Perdew, J. P. *Phys. Rev. B* **69**, 075102 (2004).
11. Schutz, M., Werner, H. J., Lindh, R. & Manby, F. R. *J. Chem. Phys.* **121**, 737–750 (2004).

Evolutionary biology

Butterfly mimics of ants

Jeremy A. Thomas and Josef Settele

Large blue butterflies are notable for their rarity and ability to dupe ants, and they are endangered. A genetic reconstruction of how social parasitism evolved among them will overturn conservation priorities.

Ants are such formidable predators that perhaps 100,000 other species of insect have evolved mechanisms to coexist with them¹. Adaptations include armour to resist attack, mimicry to avoid detection, and secretions such as honeydew to feed or appease them². In general, both partners benefit: in return for honeydew, the ants protect aphids from enemies. But natural selection can also favour cheats. It is a short evolutionary step from possessing the attributes to live safely among ants to deploying them against a colony.

Thus, among insects as diverse as butterflies, crickets, beetles and flies are specialist 'social parasites', perhaps 10,000 species in all, equipped to penetrate the highly protected chambers inside ant nests and feed, isolated

from enemies, on the rich resources concentrated there. Writing on page 386 of this issue, Als and co-workers³ provide the first molecular-genetic reconstruction of one such evolutionary pathway, that of large blue butterflies (genus *Maculinea*), including the pathway's divergence into two remarkable strategies for exploiting ants.

The large blues form a small genus that has become an icon for conservation across Europe and Asia. The adults fly in summer, laying eggs on specific plants. After two to three weeks of eating flowers, the caterpillar settles beneath its food plant to await discovery by red ants (*Myrmica*). By secreting hydrocarbons that mimic those made by *Myrmica*⁴, the caterpillar tricks a foraging worker into taking it into the nest, where it is



Figure 1 A cuckoo butterfly. This species (Rebel's large blue, *Maculinea rebeli*) survives in a few European alpine meadows. Its white eggs are laid on a particular plant, cross-leaved gentian, and the young caterpillar initially feeds on the flowers. The caterpillar then tricks a species of red ant, *Myrmica schencki*, into taking it into the ant nest. There it lives like a cuckoo (inset), being fed by nurse ants that are fooled by the insect's chemical and behavioural mimicry of the ant's own grubs, before completing its life cycle with the pupa and adult stages.

J.A. THOMAS



100 YEARS AGO

Some fifty years ago Japan was a hermit nation more than five centuries behind the times, to-day she constitutes a new and important factor in the problem of the distribution of the world's commerce. The story of the foreign commerce of Japan since the restoration of imperial authority in 1868 is told by Mr. Yukimasa Hattori... Two remarks towards the end of his paper will show the conclusions to which Mr. Hattori has come. "Japan must rely on industrial development rather than on agriculture, and must try to excel in the quality of goods produced rather than in quantity." "Japan possesses all the advantages necessary to make her a great manufacturing country. Her people possess exceptional skill, and labour is relatively cheap: coal is abundant, and the raw material is easily obtainable either at home or in the neighbouring countries." Those readers who have followed the steps in Japan's development since 1868 will be prepared to agree with Mr. Hattori that his country is but "at the very beginning of beginnings" of what will yet be seen. From *Nature* 17 November 1904.

50 YEARS AGO

Freedom, Loyalty, Dissent by Prof. Henry Steele Commager. Like Prof. A. Macbeath's plea for heretics at the British Association meeting in Belfast, the five essays collected in this volume develop the pragmatic argument that the right to dissent, even to be wrong, is an imperative necessity for a society in which science or any other creative activity is to flourish... Prof. H. S. Commager's essays are addressed to an American audience, but quite apart from the evidence they afford that there are those in the United States itself who have a national audience and are not afraid to attack McCarthyism, his refreshing defence of intellectual freedom will appeal to all scientists who welcomed Prof. Macbeath's forthright words. In the accents of Mill rather than of Burke, he argues for the encouragement of the experimental attitude in science, because "if we create a climate of opinion in which scientists fear to be bold and original or if we require that they work only on projects that appear to be of immediate importance to us, we shall end up with scientists and scientific knowledge inadequate to the tasks that we impose upon them". From *Nature* 20 November 1954.

placed among the ant grubs. In most species — the 'predatory' large blues — the caterpillar then moves to safer chambers, returning periodically to binge-feed on ant grubs. But in two 'cuckoo' species (Fig. 1), the caterpillars remain among the brood and become increasingly integrated with their society. Nurse ants feed them directly, neglecting their own brood, which may be cut up and recycled to feed the parasites⁵.

Cuckoo-feeding is an efficient way to exploit *Myrmica*, resulting in six times more butterflies per nest than is achieved by the predatory species⁶. The downside is that social acceptance is won only through secreting chemicals that so closely match the recognition codes of one host species that survival with any other ant is unlikely⁷. Thus, a typical population of a cuckoo *Maculinea* species depends exclusively on a single *Myrmica* species — which, however, differs in different regions of Europe⁸. Predatory *Maculinea* are more generalist; nevertheless, each species survives three to five times better with a single (and different) species of *Myrmica*⁶.

Theory suggests that the adaptations of cuckoo-feeders evolved from predatory ancestors, rather than vice versa⁶. It was also suspected that, because *Myrmica* ants have spread across Eurasia to form many similar-looking, physiologically distinct cryptic species, their *Maculinea* parasites, especially cuckoo-feeders, may have experienced a parallel radiation, with new species exploiting new hosts when these evolved⁸. The taxonomy of *Maculinea* has been notoriously confusing, with some authorities recognizing five species, others twelve.

Als and colleagues' molecular results³ support some ideas but overturn others. Working on seven *Maculinea* species (including all undisputed ones), and three parasitic species from their presumed sister genus *Phengaris*, they confirm that both butterfly groups evolved from a common ancestor, most probably a predatory social parasite. Cuckoo-feeding clearly evolved twice after the genera diverged, once in *Maculinea*, once in *Phengaris*. We suggest that it may have begun a third time, in the dusky large blue (*M. nausithous*); this species is still a predator, but has several attributes of cuckoo-feeders and achieves some social integration.

The surprise in the results lies in the relatedness of different species and populations. All predatory large blues contain such major differentiations that, as Als *et al.*³ point out, certain populations "may represent cryptic species". The authors do not speculate how many might exist, but a conservative interpretation of the data suggests that *M. arion*, *M. teleius*, *M. nausithous* and *M. arionides* may each be split into two similar-looking species, making (with *M. cyanecula*, here established as a true species) nine different

predatory *Maculinea* species where four had previously been recognized. A mere 21 predatory populations were sampled for this analysis, so the results raise the evolutionist's dream (but the conservationist's nightmare) that many other populations of these globally threatened butterflies may represent separate species, each more endangered than their original 'morpho-species'.

On the other hand, Als *et al.* found little evidence for the predicted cryptic speciation among cuckoo species. Indeed, they suggest that two recognized species (*M. alcon* and *M. rebeli*) might conventionally be regarded as one. Frustratingly, the ant host is known for only a few of the populations used for genetic analysis. The next challenge is to see whether genetic divergences (or similarities) in 'species' match host or other biological shifts. There are hints that some may: the genetically distinct Japanese and European populations of *M. teleius* specialize on very different *Myrmica* species, whereas the genetically close East European populations of *M. alcon* and *M. rebeli* mimic chemically similar ants and could indeed be one species.

Why is this important for conservation? The *Maculinea* have been flagship species since their selection as one of three priorities for butterfly conservation by the World Conservation Union (Queen Alexandra's birdwing of Papua New Guinea, and the Mexican roosts of monarch butterflies were the others). Their populations were undoubtedly endangered, but the problem is amplified if the recognized species consist of several cryptic forms, many needing a different ant with a different habitat requirement.

Als *et al.* rightly argue for fresh priorities. We will need to study the species status and host-use of surviving populations of predatory large blues across their entire ranges. Furthermore, the neglected genus *Phengaris* should be afforded the same priority status as *Maculinea*: it is equally specialized and endangered, and may hold the key to understanding the early evolutionary history of this intriguing group. ■

Jeremy A. Thomas is at the NERC Centre for Ecology and Hydrology's Dorset Laboratory, Winfrith Technology Centre, Dorchester DT2 8ZD, UK.
e-mail: jat@ceh.ac.uk

Josef Settele is in the Department of Community Ecology, UFZ Leipzig-Halle, Theodor-Lieser-Str. 4, 06120 Halle/Saale, Germany.
e-mail: josef.settele@ufz.de

1. Elmes, G. W. in *Biodiversity Research and its Perspectives in East Asia* (eds Lee, B. H., Kim, T. H. & Sun, B. Y.) 33–48 (Chonbuk Natl Univ., Korea, 1996).
2. Hölldobler, B. & Wilson, E. O. *The Ants* (Springer, Berlin, 1990).
3. Als, T. D. *et al. Nature* **432**, 386–390 (2004).
4. Akino, T., Knapp, J. J., Thomas, J. A. & Elmes, G. W. *Proc. R. Soc. Lond. B* **266**, 1419–1426 (1999).
5. Elmes, G. W., Wardlaw, J. C., Schönrogge, K. & Thomas, J. A. *Ent. Exp. Appl.* **110**, 53–63 (2004).
6. Thomas, J. A. & Elmes, G. W. *Ecol. Entomol.* **23**, 457–464 (1998).
7. Schönrogge, K. *et al. J. Chem. Ecol.* **30**, 91–107 (2004).
8. Elmes, G. W. *et al. Mem. Zool.* **48**, 55–68 (1994).

Nonlinear optics

Disorder is the new order

Sergey E. Skipetrov

Pure, perfectly regular crystals were believed to be essential for the efficient operation of nonlinear optical devices. Surprisingly, it now seems that disordered materials might actually perform better.

Optical frequency converters are widely used to generate coherent light at frequencies at which laser light is unavailable. Standard devices can have a very high efficiency (close to 100%), but they employ pure, sizeable — and hence expensive — nonlinear single crystals, and require careful adjustments. On page 374 of this issue, Baudrier-Raybaut *et al.*¹ show that efficient optical frequency conversion can be achieved in disordered polycrystalline materials, which are, by contrast, rather cheap to fabricate and require hardly any control. This finding is likely to mark a significant step towards large-scale applications of nonlinear optics in everyday life.

Suppose that light of some frequency (say, the red light of a ruby laser, at 4.3×10^{14} hertz) is shone onto a nonlinear crystal (a material that is not centrosymmetric). The vibrating electric field of the laser beam excites oscillations of the electrons bound in the atoms of the material. The electrons, in their turn, re-emit light at the 'fundamental' (original) frequency but also at a frequency that is double the original value (8.6×10^{14} hertz, in this example; ultraviolet light). This 'second harmonic' appears because of the anharmonicity and the asymmetry of the electric potential seen by the electrons in the crystal. This is the simplest optical frequency converter.

But nonlinear frequency conversion is efficient only if the second-harmonic waves generated by different atoms interfere constructively — or, at least, do not extinguish each other because they are out of phase (Fig. 1a). This 'phase-matching' condition is a manifestation of momentum conservation and is of paramount importance for all nonlinear wave-mixing processes. In the first report of the generation of optical harmonics by Franken *et al.*², for example, the nonlinear process was not phase-matched and the resulting signal was weak — so weak, it is said, that the editorial staff of *Physical Review Letters* mistook it for irrelevant noise and carefully removed the tiny spot it produced on the photographic plate.

The condition of phase matching requires that the two waves, the fundamental and the second harmonic, should travel with the same velocity. But the speed of a wave is a monotonically decreasing function of frequency (for so-called normal dispersion), so phase matching cannot be obtained 'for free'. The most common approach has been to

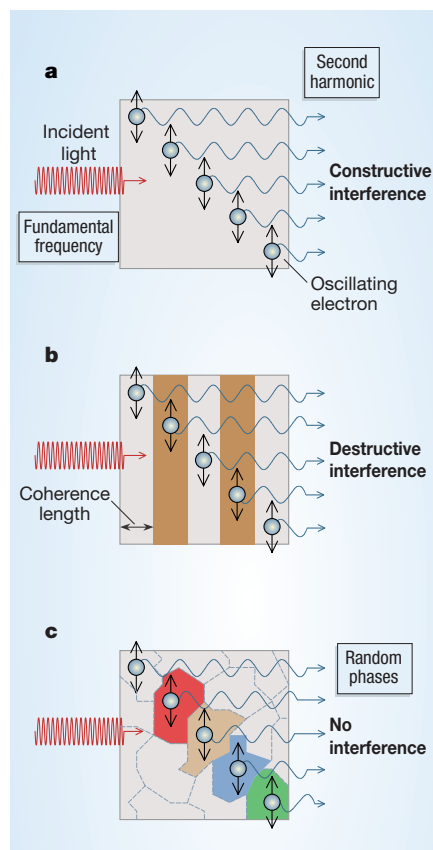


Figure 1 Phase matching. Light of a certain frequency can generate light of double the original (fundamental) frequency as it passes through a material. a, Generation of this 'second harmonic' is phase-matched when both the fundamental and the second-harmonic waves travel with the same velocity. Waves generated by electrons bound to different atoms add up in phase and the resulting intensity of the second-harmonic wave scales quadratically with the number of atoms or the length of the sample. b, In most materials, however, the two waves do not travel at the same speed and phase matching is absent. The waves generated in successive layers that are one coherence-length thick have opposite phases and interfere destructively. c, Baudrier-Raybaut *et al.*¹ show that random phase-matching can be achieved in a disordered polycrystalline material. Distinct single-crystal domains (shown in different colours) generate waves with random phases, precluding any interference effects. The resulting intensity is the sum of intensities due to individual domains and hence grows linearly with the number of domains or the length of the sample.

make use of a particular type of crystal, one that is birefringent. In such a crystal, the polarization of the light (the orientation of its electric-field vector) and its direction of propagation influence the speed of the wave. So for some special choice of polarization and direction, phase matching can be achieved.

Isotropic crystals that lack optical birefringence are a less restrictive choice, but phase matching in these is not possible: in successive layers defined by the 'coherence length' (which is related to the speed and frequency of the waves), the second-harmonic waves generated have opposite phases and interfere destructively, cancelling each other out (Fig. 1b). In a sample with an even number of coherence-length layers, the second-harmonic signal is zero; for an odd number of layers it is only as strong as if the total length of the sample were equal to the coherence length (typically 1–100 μm), because only the last layer contributes to the total signal. 'Quasi-phase matching' can be realized, however, by engineering a microstructured sample made up of layers of coherence-length width with alternating crystal orientations. The result is that the alternating sign of the nonlinear interaction makes the interference of waves generated by successive layers globally constructive. But the technology required to fabricate such a material is unfortunately rather involved, expensive and not widely accessible.

Could there be an easy, cheap solution to the phase-matching puzzle in isotropic materials? The possibility that random structures might be appropriate had already been noted^{3,4}, and now it seems that a competitive alternative to the traditional approach might indeed be provided by the 'random' quasi-phase-matching introduced by Baudrier-Raybaut *et al.*¹ (and explored theoretically as 'stochastic' quasi-phase-matching by Morozov and Chirkin⁵). These authors used a polycrystalline disordered sample, consisting of a large number of single-crystal domains with random orientations, random shapes and random sizes. The frequency-converted waves generated by different domains — be they second-harmonic waves, as in our example, or 'difference-frequency' waves from the mixing of two incoming waves, as in ref. 1 — achieve random phases and interfere neither constructively nor destructively. The total intensity of the generated wave is then the sum of the intensities arising from individual domains and it grows linearly with the number of domains or the length of the sample (Fig. 1c).

Clearly the random phase-matching is less efficient than the 'truly' phase-matched process, because the latter benefits from the constructive interference of waves generated by different parts of the nonlinear medium. But it outperforms the phase-mismatched

process for which the interference of partial waves is destructive. Random quasi-phase-matching does not require the growth of large single crystals, and needs neither a careful alignment of the optical setup nor the high-precision engineering of microstructured samples. The new technique is well suited to isotropic semiconductors (such as ZnSe, GaAs and GaP), which are technologically important and industrially mature materials. Hence, an optical frequency converter employing this technique should be substantially cheaper and more user-friendly than previously available devices. This could make the nonlinear optical technology accessible to a much wider range of potential users.

The disordered samples used by Baudrier-Raybaut *et al.*¹ have very low scattering losses — that is, they are transparent. Second-harmonic generation has also been reported in porous, strongly scattering (and hence opaque) semiconductors⁶ (in GaP), showing that disorder increases the efficiency of nonlinear frequency conversion in this case too. In fact, there is already a whole family of

wave phenomena for which disorder can be beneficial: laser action⁷, wireless communications⁸, optical cryptography⁹ and time-reversal focusing¹⁰, to name a few. Nonlinear optics now joins that list. There is obviously considerable technological potential for disordered systems — potential that is yet to be properly exploited. ■

Sergey E. Skipetrov is at the Laboratoire de Physique et Modélisation des Milieux Condensés/CNRS, 25 Avenue des Martyrs, 38042 Grenoble, France.

e-mail: sergey.skipetrov@grenoble.cnrs.fr

1. Baudrier-Raybaut, M., Haïdar, R., Kupeczek, Ph., Lemasson, Ph. & Rosencher, E. *Nature* **432**, 374–376 (2004).
2. Franken, P. A., Hill, A. E., Peters, C. W. & Weinreich, G. *Phys. Rev. Lett.* **7**, 118–119 (1961).
3. Miller, R. C. *Phys. Rev.* **134**, A1313–A1319 (1964).
4. Dewey, C. F. & Hocker, L. O. *Appl. Phys. Lett.* **26**, 442–444 (1975).
5. Morozov, E. Yu. & Chirkin, A. S. *Quantum Electronics* **34**, 227–232 (2004).
6. Mel'nikov, V. A. *et al. Appl. Phys. B* **79**, 225–228 (2004).
7. Wiersma, D. S. *Nature* **406**, 132–134 (2001).
8. Simon, S. H., Moustakas, A. L., Stoytchev, M. & Safar, H. *Phys. Today* **54** (9), 38–43 (2001).
9. Pappu, R., Recht, B., Taylor, J. & Gershenfeld, N. *Science* **297**, 2026–2030 (2002).
10. Fink, M. *Phys. Today* **50** (3), 34–40 (1997).

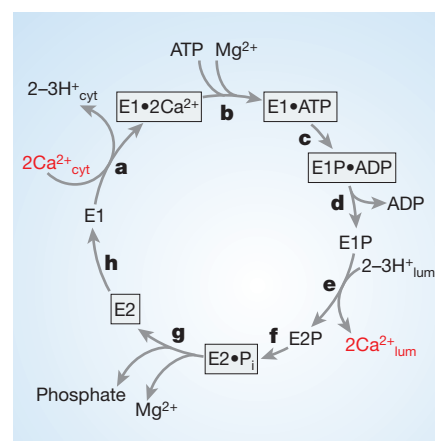


Figure 1 The Post-Albers cycle: states of the ion pump Ca^{2+} -ATPase as it transports Ca^{2+} ions into the sarcoplasmic reticulum. Only the forward direction is indicated, although the cycle can run in reverse under favourable conditions *in vitro*. a, A rise in the Ca^{2+} concentration in the cytoplasm leads to the binding of two cytoplasmic Ca^{2+} ions ($\text{Ca}^{2+}_{\text{cyt}}$) to the pump's two high-affinity Ca^{2+} -binding sites, forming $\text{E1}\cdot 2\text{Ca}^{2+}$. b, ATP (plus Mg^{2+}) binds. c, ATP is hydrolysed, producing ADP plus phosphate; the phosphate is used by Ca^{2+} -ATPase to phosphorylate itself on a key aspartate amino-acid residue. d, ADP is released. By now, conformational changes have ensured that the Ca^{2+} ions are occluded within the protein and so cannot access the cytoplasm. e, In a rate-limiting step to generate E2P , the enzyme loses its ability to rephosphorylate ADP and its high affinity for Ca^{2+} , and opens its gate to the interior (lumen) of the sarcoplasmic reticulum, releasing the Ca^{2+} ions ($\text{Ca}^{2+}_{\text{lum}}$). f, Water (not shown) enters the catalytic site and hydrolyses the phosphorylated aspartate. g, After release of phosphate (plus Mg^{2+}), the ATPase (E2) is regenerated. E2 has low-affinity Ca^{2+} -binding sites exposed to the lumen; Ca^{2+} binding from the cytoplasmic side requires conversion to E1 (h). The five states relevant to Fig. 2 are outlined.

Structural biology

Ion pump in the movies

C. Roy D. Lancaster

Insight into how membrane ion pumps work requires structural snapshots of various stages of their catalytic cycle. Now a fifth freeze-frame image of a calcium pump in action adds to a striking body of work on this protein.

On page 361 of this issue, Toyoshima and colleagues¹ reveal the latest in a series of high-resolution, three-dimensional snapshots of an ion pump at work. The pump in question is powered by adenosine triphosphate (ATP) molecules and moves calcium ions across biological membranes — an activity that is crucial for skeletal muscle to function. Comparison of this structural snapshot with images of other stages of the pump's catalytic cycle provides an almost-complete movie of its mechanism.

Discovered more than 40 years ago in rabbit muscles^{2,3}, the subject of Toyoshima and colleagues' studies is known as Ca^{2+} -ATPase. It is found in the membrane of a storage compartment, the sarcoplasmic reticulum, in muscle cells, and functions to pump Ca^{2+} ions into this compartment, causing the muscles to relax. It is also representative of a broader superfamily of proteins known as P-type ATPases (reviewed in refs 4, 5), so named because their activity is driven by the hydrolysis of ATP and the reaction is associated with the covalent attachment of a phosphate group to a particular aspartate amino-acid residue in the proteins themselves, forming high-energy intermediates. Other members of this superfamily include the Na^+ , K^+ -ATPase, which

controls ion balance and membrane potential in all animal cells, and the gastric H^+ , K^+ -ATPase, a target for drugs that treat gastro-oesophageal acid reflux. So an understanding of how the Ca^{2+} -ATPase pumps ions should illuminate the general mechanism of action of these proteins too.

Over the years, the inner workings of the Ca^{2+} -ATPase have gradually been elucidated. Functionally, it is known that, for each molecule of ATP hydrolysed, the Ca^{2+} -ATPase transports two Ca^{2+} ions into the interior (the lumen) of the sarcoplasmic reticulum, and moves two to three H^+ ions in the opposite direction⁶. According to a popular model known as the Post-Albers cycle^{7,8}, such transport is achieved by changing the affinity and accessibility of the Ca^{2+} -binding sites of Ca^{2+} -ATPase. Thus, these sites switch between high-affinity states (known as E1) that face the cytoplasm — outside the sarcoplasmic reticulum — and low-affinity states (E2) that are open to the lumen⁹. A simplified scheme of this cycle (Fig. 1) depicts the various distinct conformational forms of the E1 and E2 states.

But, mechanistically, how exactly does this ion pump pump? Over the past four years Toyoshima and colleagues have led the crusade to understand its mechanism,

painstakingly producing three-dimensional structural snapshots of the Ca^{2+} -ATPase at various stages of its catalytic cycle (Fig. 2). First they reported¹⁰, at 2.6 Å resolution, the crystal structure of the ATP-free, Ca^{2+} -bound E1 state, designated $\text{E1}\cdot 2\text{Ca}^{2+}$, which was a prerequisite for understanding the detailed mechanism of this Ca^{2+} pump. This structure also provided the first atomic model for any P-type ATP-dependent cation pump. It revealed that, in this state, two Ca^{2+} ions are bound in the protein's trans-membrane region, which consists of ten structural features known as α -helices. The cytoplasmic part of the pump, meanwhile, was found to consist of three regions, or domains. The phosphorylation (P) and nucleotide-binding (N) domains form the catalytic site, in which phosphorylation and ATP hydrolysis occur, whereas the actuator

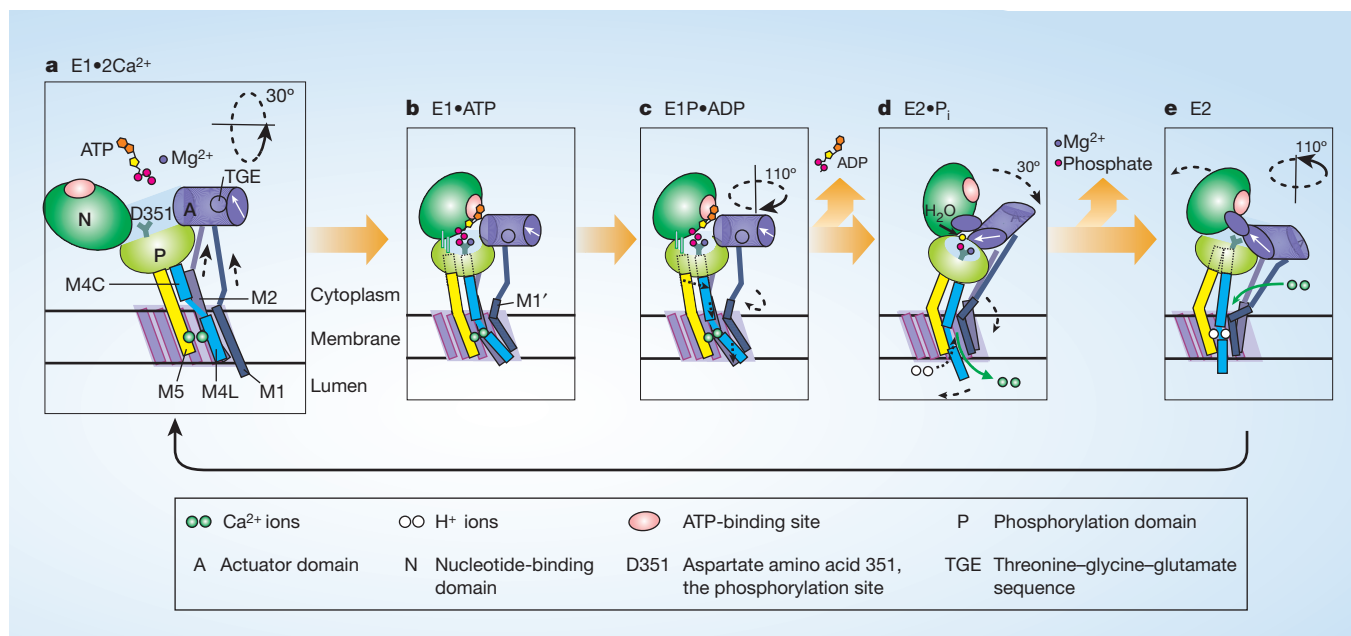


Figure 2 Ca^{2+} -ATPase in the five key states for which high-resolution structures are available. **a**, Two Ca^{2+} ions are bound¹⁰ in the high-affinity sites formed by transmembrane helices M4, M5, M6 and M8, M6 and M8 being behind M5 and M4. (M4 has two parts, cytoplasmic, C, and luminal, L.) Helix M1 is embedded in the membrane. The cytoplasmic gate is open. Arrows denote movements that occur to produce the next state. **b**, ATP binds and crosslinks the P- and N-domains, causing the P-domain to bend¹³. The N-domain is fixed in a highly inclined position and contacts the A-domain in a strained position. M1 is pulled towards the cytoplasm and bent, so that its top (M1') closes the cytoplasmic gate. **c**, **d**, ATP hydrolysis

releases a phosphate group for transfer to D351; ADP dissociates, opening the interface between the N- and P-domains. The A-domain rotates so that its TGE loop wedges into the gap and interacts with D351. Large movements of M4 towards the lumen, sharp bending of M5 and rotation of M6 destroy the Ca^{2+} -binding sites¹¹. Sections of M1 and M2 push against M4L, opening the luminal gate and releasing Ca^{2+} . The TGE loop fixes a water molecule and catalyses its attack on the aspartate phosphate¹. **e**, Release of phosphate and Mg^{2+} straightens the P-domain, releasing M1 and M2 so that M4L can close the luminal gate¹. Conversion of E2 to E1 will straighten M5 and break the closed configuration of the N-, P- and A-domains.

(A) domain, formerly called the transducer or beta domain, seems to be involved in transmitting major conformational changes throughout the protein.

Two years ago, Toyoshima and Nomura¹¹ provided another milestone in understanding the mechanism of Ca^{2+} -ATPase — the structure of the pump in its Ca^{2+} -free E2 state, at 3.1 Å resolution. This structure was stabilized by a potent inhibitor of the enzyme, thapsigargin, which locks it in a form analogous to E2 (ref. 12). Then, earlier this year, Toyoshima and Mizutani¹³ and Sørensen and colleagues¹⁴ independently contributed structures of Ca^{2+} -ATPase with the bound ATP analogue AMPPCP — structures thought to mimic that of the ATP- and Ca^{2+} -bound E1 state, E1•ATP.

Sørensen and colleagues¹⁴ also provided a structure of the pump with the bound phosphate analogue aluminium fluoride, probably resembling the E1P•ADP state — the E1 state in which two Ca^{2+} ions are bound, ATP has been hydrolysed (producing, but not yet releasing, adenosine diphosphate, ADP), and the pump is phosphorylated on the key aspartate. Toyoshima and colleagues¹ now present their own structure of the E1P•ADP state. They also describe, at 2.3 Å resolution, the structure of a Ca^{2+} -free, thapsigargin-stabilized Ca^{2+} -ATPase with the bound phosphate analogue magnesium fluoride. This structure is thought to approximate a

fifth intermediate state in the Post-Albers cycle, E2•P_i — a form in which ADP has been released and in which the phosphate group attached to the aspartate has been hydrolysed but not yet released. On the basis of these latter structures, the authors conclude that the release of ADP triggers the opening of the so-called luminal gate, allowing the bound Ca^{2+} ions access to the interior of the sarcoplasmic reticulum. The later release of phosphate, meanwhile, must trigger the closure of this luminal gate.

With all of these crystal structures to hand, representing five different intermediate states, Toyoshima and colleagues have for the first time deduced an almost complete movie of the structural changes associated with ion pumping by a P-type ATPase (Fig. 2; see also the supplementary movie associated with ref. 1). In essence, the P- and N-domains change interfaces and thereby control the position of the A-domain, which in turn transmits information to the transmembrane Ca^{2+} gates. ATP, phosphate, Mg^{2+} and Ca^{2+} are the modifiers of these interfaces.

Of course, the structures solved by Toyoshima and colleagues, and by others, are invariably equilibrium approximations of transient intermediate states. Consequently, speculation remains regarding the precise structural details of the true intermediate states. However, the large structural differences observed between four of the five

structures — and the fact that, as Toyoshima and colleagues¹ discuss, these differences and functional studies have pinpointed the same amino acids as being essential to the pump's action — reassure us that these structural studies have captured the essence of the mechanism of this membrane protein in a hitherto unrivalled manner.

C. Roy D. Lancaster is in the Department of Molecular Membrane Biology, Max Planck Institute of Biophysics, Marie-Curie-Strasse 15, D-60439 Frankfurt am Main, Germany.
e-mail: roy.lancaster@mpibp-frankfurt.mpg.de

1. Toyoshima, C., Nomura, H. & Tsuda, T. *Nature* **432**, 361–368 (2004).
2. Hasselbach, W. & Makinose, M. *Biochem. Z.* **333**, 518–528 (1961).
3. Ebashi, S. & Lipman, F. J. *Cell Biol.* **14**, 389–400 (1962).
4. Möller, J. V., Juul, B. & le Maire, M. *Biochim. Biophys. Acta* **1286**, 1–51 (1996).
5. Kühlbrandt, W. *Nature Rev. Mol. Cell. Biol.* **5**, 282–295 (2004).
6. Yu, X., Carroll, S., Rigaud, J. L. & Inesi, G. *Biophys. J.* **64**, 1232–1242 (1993).
7. Post, R. L., Hegyváry, C. & Kume, S. *J. Biol. Chem.* **247**, 6530–6540 (1972).
8. Albers, R. W. *Annu. Rev. Biochem.* **36**, 727–756 (1967).
9. de Meis, L. & Vianna, A. L. *Annu. Rev. Biochem.* **48**, 275–292 (1979).
10. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. *Nature* **405**, 647–655 (2000).
11. Toyoshima, C. & Nomura, H. *Nature* **418**, 605–611 (2002).
12. Sagara, Y. & Inesi, G. *J. Biol. Chem.* **266**, 13503–13506 (1991).
13. Toyoshima, C. & Mizutani, T. *Nature* **430**, 529–535 (2004).
14. Sørensen, T. L., Möller, J. V. & Nissen, P. *Science* **304**, 1672–1675 (2004).

Chemical biology

Light switch for proteins

J. Am. Chem. Soc. **126**, 14306–14307 (2004)

An enzyme has been given light-switchable activity by having one of its key amino acids replaced with a 'photocaged' version.

Ning Wu and colleagues have accomplished this feat in yeast by developing a way of loading transfer RNA (tRNA) with the non-natural amino acid. The tRNA molecule carries the amino acid to the cell's protein-synthesizing machinery, the ribosome, where it is stitched into the growing protein chain. This approach has been used previously to incorporate many non-natural amino acids into proteins. The trick is twofold. First, an artificial codon, or coding unit of DNA, is built into the gene encoding the protein, and this is recognized in the messenger RNA transcript only by the tRNA loaded with the non-natural amino acid. Second, the enzyme responsible for 'charging' the tRNA must be modified to accept that new amino acid.

Using *in vitro* selection, Wu *et al.* found mutant enzymes that attach three non-natural amino acids to a tRNA that normally carries leucine. That way, they replaced a cysteine residue in the active site of the protease enzyme caspase 3 with *o*-nitrobenzyl cysteine. Ultraviolet light removes the nitrobenzyl group and triggers enzyme activity.

Philip Ball

Structural biology

Some like it (very) cold

Proc. Natl Acad. Sci. USA
doi:10.1073/pnas.0405109101 (2004)

Although X-ray crystallography has been used to elucidate the molecular mechanisms of many biological processes, it rarely reveals the exact position of important hydrogen atoms. Neutron crystallography, by contrast, can determine the location of these atoms. But collecting data at low temperatures, which often increases the number of ordered residues, is a technical challenge, because the crystals required for neutron crystallography are quite large.

Recently, however, it has proved possible to cool large crystals of the sugar-binding protein concanavalin A to 15 K. Now, M. P. Blakeley *et al.* report the neutron structure of the saccharide-free protein at 2.5 Å resolution. This structure contains twice as many ordered water molecules as the authors' previously published structure at 293 K (room temperature). In that structure, the nuclear density for water

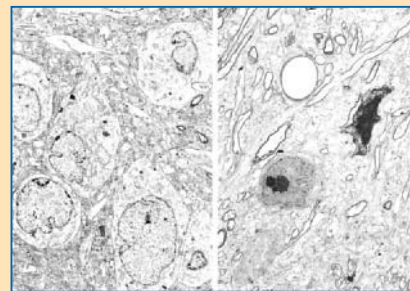
Neurodegenerative diseases

Model matters

Hum. Mol. Genet. doi:10.1093/hmg/ddi004 (2004)

Huntington's disease remains a devastating neurodegenerative disorder for which no cure exists, although scientists continue to investigate its genetic and cellular causes. Their understanding of the disease might benefit from an improved experimental model: the most common mouse model reproduces many of the clinical features, such as impaired coordination, but, strangely, little neuronal death has hitherto been detected in these mice.

Åsa Petersén *et al.* now show, however, that the animals do suffer a dramatic deterioration and loss of certain brain cells, as shown here — the dark spots on the right image represent degenerating nerve cell bodies; the left image is from a normal mouse. The dying neurons are those in the lateral hypothalamus that express the peptide orexin (also called hypocretin). More specifically, mice in the final stages of the disease had 72%



fewer orexin-expressing neurons than controls — and the same percentage decrease in orexin levels in cerebrospinal fluid. The animals were also narcoleptic (prone to excessive daytime sleepiness), as are people with impaired orexin function.

The authors found similarly affected orexin-expressing neurons in patients with Huntington's disease. So Petersén *et al.* suggest that the loss of orexin could be used to assess the progression of the illness.

Roxanne Khamis

molecules in the saccharide-binding site was poorly defined, but at 15 K there is an extensive hydrogen-bonding network, involving five water molecules and their interacting amino acids.

Low-temperature neutron structures of other proteins have been solved (the authors refer to unpublished data for lysozyme and rubredoxin). Analysis of these structures at 15 K, and comparison with room-temperature structures, could teach us a great deal about protein function.

The development also opens up the possibility of time-resolved, neutron protein crystallography via freeze-trapping.

Joshua Finkelstein

Physics

In a spin

Science doi:10.1126/science.1105514 (2004)

When an electron travels at right angles to the direction of a magnetic field, it experiences a force along a third axis that is perpendicular to both. This 'Hall effect' makes a stream of electrons veer off course, and can produce a build-up of charge on one side of an electrical conductor.

Y. K. Kato *et al.* now present the first experimental evidence for the 'spin Hall effect', which is due to the magnetic field generated by the angular momentum of the electrons themselves, and is predicted by theory. This leads to 'spin up' and 'spin down' electrons moving in opposite directions transverse to an applied electric field, even in the absence of an external magnetic field: spin currents without net charge flow. The authors' observations of gallium arsenide and indium gallium arsenide semiconductors show that one side

of the material accumulates electrons that spin in one direction, while the other side becomes enriched with electrons spinning in the opposite direction — just as predicted.

The spin Hall effect could now be used to manipulate electron spins without using magnetic fields, the authors note. Potential applications lie in the emerging field of spin electronics.

Mark Peplow

Materials chemistry

Protean paper

Adv. Mater. **16**, 1729–1732 (2004)

Filters, catalysts and sensors are some of the possible applications of a nanostructured 'paper' material made by Jikang Yuan and colleagues. The material consists of fibres of manganese oxide arranged into stacks of sheet-like aggregates, and it is — given its high degree of hierarchical ordering — astonishingly easy to make.

Each individual fibre is crystalline, with a molecular-sieve structure perforated by linear channels just a few ångströms wide. This microporous phase of manganese oxide is known already, and is made by heating a mixture of metal salts in water. The porous fibres that result are about 30 nm wide and several micrometres long, and they precipitate to form a tangled, woolly mass. When this is coated onto a solid substrate and heated, it produces a papery film that can be unpeeled from the surface.

The researchers find that the fibres become organized into layers, and are oriented at right angles in successive layers. The paper can be folded and cut, and may be 'recycled' by re-dispersing the fibres in water. Metal particles deposited in the micropores could turn this substance into a membrane for size-selective catalysis.

Philip Ball

Grape ripening as a past climate indicator

Summer temperature variations are reconstructed from harvest dates since 1370.

French records of grape-harvest dates in Burgundy were used to reconstruct spring–summer temperatures from 1370 to 2003 using a process-based phenology model developed for the Pinot Noir grape. Our results reveal that temperatures as high as those reached in the 1990s have occurred several times in Burgundy since 1370. However, the summer of 2003 appears to have been extraordinary, with temperatures that were probably higher than in any other year since 1370.

Biological and documentary proxy records have been widely used to reconstruct temperature variations to assess the exceptional character of recent climate fluctuations^{1–3}. Grape-harvest dates, which are tightly related to temperature, have been recorded locally for centuries in many European countries. These dates may therefore provide one of the longest uninterrupted

series of regional temperature anomalies (highs and lows) without chronological uncertainties¹.

In Burgundy, these officially decreed dates have been carefully registered in parish and municipal archives since at least the early thirteenth century. We used a corrected and updated harvest-date series⁴ from Burgundy, covering the years from 1370 to 2003, to reconstruct spring–summer temperature anomalies that had occurred in eastern France. To convert historical observations into temperature anomalies, we used a process-based phenology model for Pinot Noir, the main variety of grape that has been continuously grown in Burgundy since at least the fourteenth century⁵ (for details, see supplementary information).

Our yearly reconstruction is significantly correlated (Table 1) with summer temperatures deduced from tree rings in central



A fifteenth-century depiction of the grape harvest from *Les Très Riches Heures du Duc de Berry*, a medieval book of hours.

France⁶ (correlation coefficient, $r=0.53$), the Burgundy part of a spatial multi-proxy reconstruction² ($r=0.69$) and observed summer temperatures in Paris⁷ ($r=0.75$), central England⁸ ($r=0.53$) and the Alps⁹ ($r=0.45$).

Figure 1 shows two early warm decadal fluctuations: one in the 1380s ($+0.72\text{ }^{\circ}\text{C}$) and one in the 1420s ($+0.57\text{ }^{\circ}\text{C}$), both above the 95th percentile. The warm period of the 1420s was followed by a cold period that lasted from the mid-1430s to the end of the 1450s ($-0.45\text{ }^{\circ}\text{C}$, under the 10th percentile). Our series also reveals particularly warm events, above the 90th percentile, in the 1520s and between the 1630s and the 1680s. These decades were as warm as the end of the twentieth century. The high-temperature event of 1680 was followed by a cooling, which culminated in the 1750s (under the 5th percentile) — the start of a long cool period that lasted until the 1970s.

The inferred anomaly for the summer of 2003 represents an unprecedented event. It was $+5.86\text{ }^{\circ}\text{C}$ warmer than the reference period (1960–89), whereas the next highest anomaly during the whole period was $+4.10\text{ }^{\circ}\text{C}$ in 1523. This confirms and refines the conclusions of previous studies^{2,10} about the exceptional warmth of the 2003 summer in France.

Grape-harvest dates offer the potential to trace geographical variations in temperature over large parts of Europe and the Middle East over past centuries. This climate, history and phenology synergy can be used to reconstruct temperatures that will substantially add to the long proxy-record databases and provide insight into regional-scale climate variations.

Isabelle Chuine*, Pascal Yiou†, Nicolas Viovy†, Bernard Seguin‡, Valérie Daux†, Emmanuel Le Roy Ladurie§

Table 1 Linear correlation coefficients between reconstructed temperatures

Series	Tree rings (JJA)	Multi-proxy (JJA)	Multi-proxy (AMJJA)	Paris (AMJJA)	Central England (AMJJA)	Alps (JJA)
Time range	1750–1975	1500–1998	1659–1998	1787–2000	1663–1992	1760–1998
Linear correlation coefficient, r (grape-harvest date)	0.53 ± 0.09	0.57 ± 0.06	0.69 ± 0.06	0.75 ± 0.07	0.53 ± 0.08	0.45 ± 0.09

Linear correlation coefficients between temperatures reconstructed from grape-harvest dates in Burgundy and other observed or reconstructed temperatures are shown. Comparison with temperatures reconstructed from a tree-ring database⁶ used the closest grid point to Burgundy in the data set. Comparison with multi-proxy reconstructed temperatures² used the four closest grid points to Burgundy in the data set. The Paris⁷, central England⁸, and Alps⁹ series are observed temperatures (instrumental series). Correlations were computed on the common time intervals between two time series. 95% confidence intervals are shown. JJA, June to August; AMJJA, April to August.

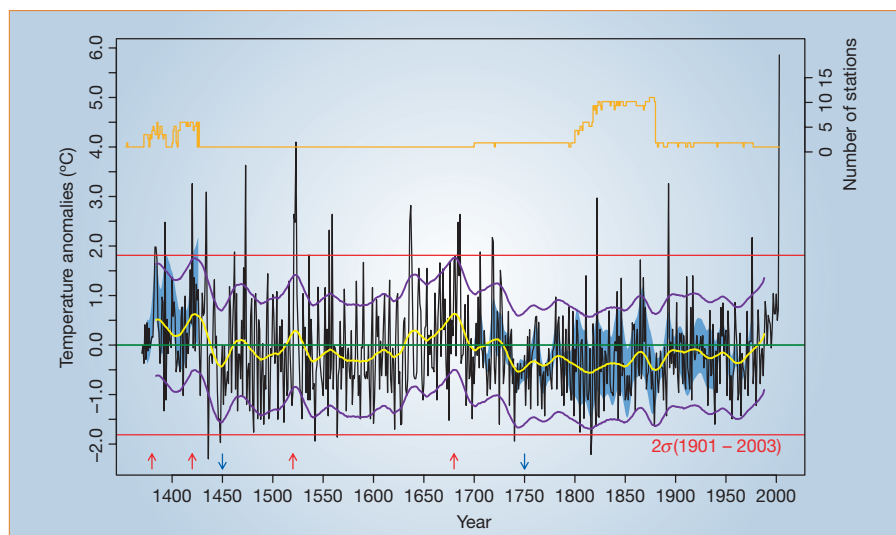


Figure 1 April–August temperature anomalies in Burgundy, France, as reconstructed from grape-harvest dates from 1370 to 2003. Yearly anomalies are in black and the 30-year gaussian filter is in yellow. Confidence intervals due to vineyard differences, with an 11-year smoothing, are shaded in blue; these are estimated from the inter-station variability upper 90th and lower 10th percentiles, and are determined when there are more than three available observations in a year. Orange line (number of stations) represents the number of observed harvest dates for each year, indicating where the confidence intervals are computed. Confidence intervals with two s.e., due to the regression between observed and reconstructed temperature in Dijon, are in purple. These were obtained by regressing the reconstructed temperature with the observed temperature over 1880–2000. Green horizontal (zero) line is determined from the 1960–89 reference period. Red horizontal lines represent the 2σ interval of the reconstructed temperature for the twentieth century (1901–2003). Vertical arrows indicate warm decadal periods (red) above the 90th percentile and the cold trends (blue) under the 10th percentile.

*CEFE-CNRS, 1919 route de Mende, 34293

Montpellier, France

†LSCE-CEA-CNRS CE Saclay l'Orme des

Merisiers, 91191 Gif-sur-Yvette, France

e-mail: pascal.yiou@cea.fr

‡INRA Site Agroparc, domaine Saint-Paul, 84914

Avignon Cedex 9, France

§Collège de France, 75231 Paris Cedex 05, France

1. Pfister, C. *Wetternachhersage. 500 Jahre Klimavariationen und Naturkatastrophen 1496–1995* (Haupt, Bern, Stuttgart and Wien, 1999).
2. Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M. & Wanner, H. *Science* **303**, 1499–1503 (2004).
3. Jones, P. D. & Mann, M. E. *Rev. Geophys.* **42**, doi:10.1029/2003RG000143 (2004).
4. Le Roy Ladurie, E. *Histoire du Climat depuis l'An Mil* (Champs Flammarion, Paris, 1983).
5. Robinson, J., Dinsmoor, A. & Smart, R. E. *The Oxford Companion to Wine* (Oxford University Press, 1999).
6. Briffa, K. R., Jones, P. D. & Schweingruber, F. H. *Quat. Res.* **30**, 36–52 (1988).
7. Renou, E. *Ann. Bur. Centr. Météorol.* **B 195**–226 (1887).
8. Manley, G. Q. *J. R. Meteorol. Soc.* **100**, 389–405 (1974).
9. Boehm, R. et al. *Int. J. Climatol.* **21**, 1779–1801 (2001).
10. Schär, C. et al. *Nature* **427**, 332–336 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Climate

Large-scale warming is not urban

Controversy has persisted^{1,2} over the influence of urban warming on reported large-scale surface-air temperature trends. Urban heat islands occur mainly at night and are reduced in windy conditions³. Here we show that, globally, temperatures over land have risen as much on windy nights as on calm nights, indicating that the observed overall warming is not a consequence of urban development.

Observations of the minimum temperature ($T_{\min}$) over 24 hours at 264 stations worldwide since 1950 were expressed as anomalies, relative to the period 1961–90 where possible. Coverage of $T_{\min}$ data was good north of 20° N, in Australasia and in the western tropical Pacific, but poor in Africa, South America, Antarctica and parts of southern Asia. Reanalysed⁴ daily-average near-surface wind components were used to classify the $T_{\min}$ anomalies into 'windy' (upper tercile) and 'calm' (lower tercile) conditions. Daily average wind speeds were used because the timings of temperature extremes are not known. For stations between 140° E and the dateline, $T_{\min}$ — which occurs most frequently in the early morning — was matched with the previous day's speed. This is because the early morning in terms of universal time (equivalent to Greenwich Mean Time) is still in the previous day in the Far East.

Annual and seasonal anomalies of $T_{\min}$ were gridded on a 5° × 5° resolution for windy, calm and 'all' conditions. Coverage was at least 200 grid boxes (equivalent to

more than 27% of global land area) in 1958–99. For 1950–2000, the trends of global annual average $T_{\min}$ for windy, calm and all conditions were identical (0.19 ± 0.06 °C per decade; Fig. 1a). So, urbanization has not systematically exaggerated the observed global warming trends in $T_{\min}$. The same can be said for poor instrumental exposure and microclimatic effects, which are also reduced when instruments are well ventilated⁵.

When the criterion for 'calm' was changed to the lightest decile of wind strength, the global trend in $T_{\min}$ was unchanged. The analysis is therefore robust to the criterion for 'calm'. To assess the effect of time differences between the reanalysis⁴ daily-average winds and $T_{\min}$, I repeated the analysis using 26 stations in North America and Siberia that have hourly or six-hourly reports of simultaneous temperature and wind. Again, windy and calm nights warmed at the same rate, in this case by 0.20 °C per decade.

Because a small sample was used, I compared global trends for the reduced period 1950–93 with published all-conditions trends for that period based on a sample of over 5,000 stations⁶. All differences were within ± 0.02 °C per decade. This robustness arises because of the spatial coherence of surface temperature variations and trends⁷.

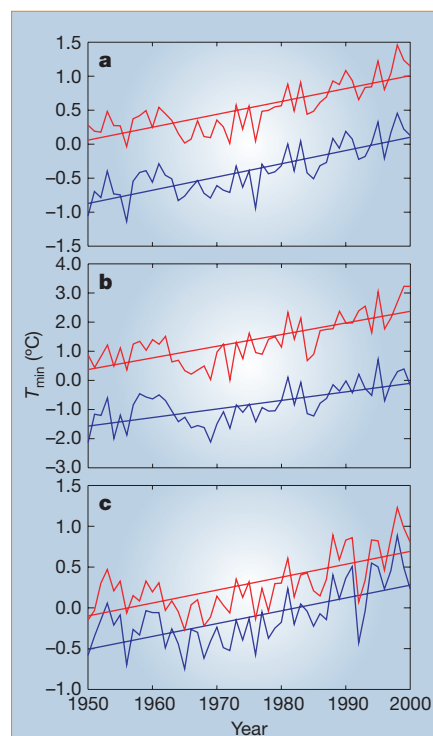


Figure 1 Anomalies in $T_{\min}$ for windy (red) and calm (blue) conditions. **a**, Annual global data; **b**, winter data (December to February) for Northern Hemisphere land north of 20° N; **c**, summer data (June to August) for Northern Hemisphere land north of 20° N. The linear trend fits, and the $\pm 2\sigma$ error ranges given in the text, were estimated by restricted maximum likelihood¹⁰, taking into account autocorrelation in the residuals. As expected from the reduced stratification of the boundary layer, $T_{\min}$ is, on average, warmer on windy nights than on calm nights.

The global annual result conceals a relative warming of windy nights in winter in the extratropical Northern Hemisphere (Fig. 1b), mainly in western Eurasia. The observed tendency to an increased positive phase of the North Atlantic Oscillation⁸ implies that the windier days in western Eurasia had increased warm advection from the ocean⁹, yielding greater warming. In summer in the extratropical Northern Hemisphere (Fig. 1c), there was no relative change in $T_{\min}$ on windy nights. At that time of year, atmospheric circulation changes are less influential, but an urban warming signal is still absent. In the tropics, calm nights warmed relative to windy nights on an annual average, but only by 0.02 ± 0.01 °C per decade, which is much less than the overall tropical warming in $T_{\min}$ (0.16 ± 0.03 °C).

This analysis demonstrates that urban warming has not introduced significant biases into estimates of recent global warming. The reality and magnitude of global-scale warming is supported by the near-equality of temperature trends on windy nights with trends based on all data.

David E. Parker

Hadley Centre, Meteorological Office,

Exeter EX1 3PB, UK

e-mail: david.parker@metoffice.com

1. Kalnay, E. & Cai, M. *Nature* **423**, 528–531 (2003).
2. Peterson, T. C. *J. Clim.* **16**, 2941–2959 (2003).
3. Johnson, G. T. et al. *Bound. Layer Meteorol.* **56**, 275–294 (1991).
4. Kalnay, E. et al. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
5. Parker, D. E. *Int. J. Climatol.* **14**, 1–31 (1994).
6. Easterling, D. R. et al. *Science* **277**, 364–367 (1997).
7. Jones, P. D., Osborn, T. J. & Briffa, K. R. *J. Clim.* **10**, 2548–2568 (1997).
8. Folland, C. K. et al. in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Houghton, J. T. et al.) 99–181 (Cambridge Univ. Press, Cambridge, UK, 2001).
9. Hurrell, J. W. & van Loon, H. *Climat. Change* **36**, 301–326 (1997).
10. Diggle, P. J., Liang, K. Y. & Zeger, S. L. *Analysis of Longitudinal Data* (Clarendon, Oxford, 1999).

Competing financial interests: declared none.

Atmospheric science

Early peak in Antarctic oscillation index

The principal extratropical atmospheric circulation mode in the Southern Hemisphere, the Antarctic oscillation (or Southern Hemisphere annular mode), represents fluctuations in the strength of the circumpolar vortex and has shown a trend towards a positive index in austral summer in recent decades, which has been linked to stratospheric ozone depletion^{1,2} and to increased atmospheric greenhouse-gas concentrations^{3,4}. Here we reconstruct the austral summer (December–January) Antarctic oscillation index from sea-level pressure measurements over the twentieth century⁵ and find that large positive values, and positive trends of a similar magnitude

to those of past decades, also occurred around 1960, and that strong negative trends occurred afterwards. This positive Antarctic oscillation index and large positive trend during a period before ozone-depleting chemicals were released into the atmosphere and before marked anthropogenic warming, together with the later negative trend, indicate that natural forcing factors or internal mechanisms in the climate system must also strongly influence the state of the Antarctic oscillation.

Until recently⁶, it has not been possible to put the Antarctic oscillation index (AAOI) trends in past decades into a longer-term context, as comprehensive Southern Hemisphere data are limited to the reanalysis period (1948/58–present; NCAR–NCEP/ERA40 reanalysis). Our reconstructions are intended to cover the reanalysis period with a consistent estimate of the AAOI, as this has been questioned⁷, and to extend this estimate further back. The new reconstructions are more reliable as they use more predictor stations and a statistical model fitted using ERA40 reanalysis, whose AAOI estimates are better than those from NCEP reanalysis⁷.

We define the Antarctic oscillation as the first empirical orthogonal function, and the AAOI as the first principal component of the December–January mean extratropical sea-level pressure. A positive or negative AAOI indicates a strengthening or weakening, respectively, of circumpolar westerly flow. For our reconstructions, we used multiple regression to estimate the AAOI from the leading principal components of normalized station pressure. The model is fitted using detrended data, but the reconstruction is derived using undetrended data. One reconstruction (1905–2000) uses 22 stations (Fig. 1a); the second (1951–2000) uses 41 and provides improved coverage of the Antarctic oscillation centres of action (Fig. 1b). Cross-validation gives a correlation of 0.88 and 0.90 for the 1905 and 1951 reconstructions, respectively. (For methods, see supplementary information).

Both reconstructions show that the current positive values for the AAOI are not unprecedented (Fig. 1c). After the relatively stable first half of the twentieth century, there is a period of positive values (relative to the 1958–2000 mean) from 1958 to 1963, followed by a sharp drop to predominantly negative values until the mid-1980s, and then by a mostly positive phase up to the present. The maximum positive 25-year trends over recent years are of similar magnitude to those between the low values of the 1940s and the peak in the 1960s. Note that the trend over the past decades is caused by a combination of negative values in the 1970s and current positive values.

A positive AAOI around 1960, followed by a negative index, is also present in the NCEP and the ERA40 data, in a zonal index-

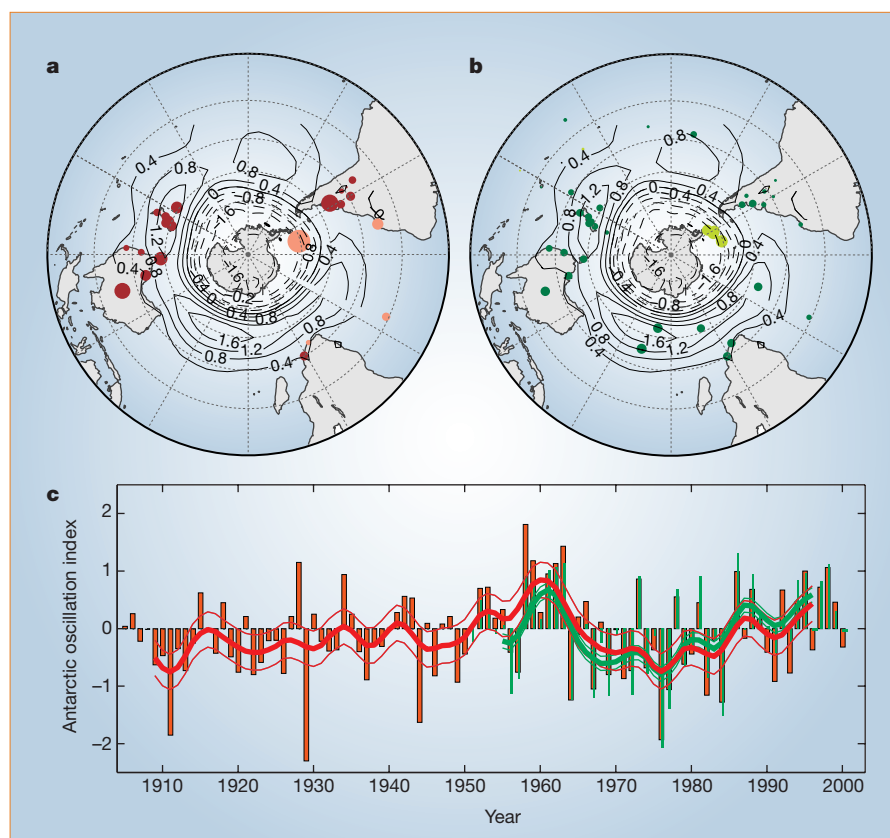


Figure 1 Reconstruction of the December–January Antarctic oscillation index (AAOI). **a**, The Antarctic oscillation pattern and regression weights for normalized station sea-level pressure used for the 1905 AAOI reconstruction. Isolines show the sea-level pressure anomaly (in hundreds of pascals) for the AAOI + 1. The red circles denote positive values and the pink circles denote negative ones; the area is proportional to the weight; **b**, as in **a**, but for the 1951 AAOI reconstruction, with dark green denoting positive values and light green denoting negative ones. **c**, Reconstructed December–January AAOI. Red bars show the 1905 reconstruction; green bars, the 1951 reconstruction. The thick red line is the nine-year low-pass-filtered 1905 reconstruction; the green, the 1951 reconstruction. The thin red and green lines show the 95% confidence intervals for the filtered data. Years are dated from December.

based AAOI⁷ and in earlier reconstructions⁶. Consistent with this Antarctic oscillation behaviour, station pressures around 1960 have positive anomalies in the mid-latitude centres of action and negative anomalies in the Antarctic centre of action. By contrast with our reconstructions, the 1960s peak is slightly lower than the 1990s peak in both reanalyses and the zonal index AAOI. Despite this small uncertainty about the exact values, the 1960s peak is a robust feature in all these data sets.

The fact that the austral summer behaviour of the Antarctic oscillation in recent decades seems not to be unprecedented indicates that natural forcing factors, such as solar or volcanic variability, or internal processes in the climate system, can strongly influence the state of the Antarctic oscillation. The question arises as to what the role of these factors has been over the past decades.

Julie M. Jones, Martin Widmann

*Institute for Coastal Research, GKSS Research Centre, 21502 Geesthacht, Germany
e-mail: jones@gkss.de*

- Thompson, D. J. & Solomon, S. *Science* **296**, 895–899 (2002).
- Gillett, N. P. & Thompson, D. W. J. *Science* **302**, 273–275 (2003).
- Stone, D. A., Weaver, A. J. & Stouffer, R. J. *J. Clim.* **14**, 3551–3565 (2001).
- Kushner, P. J. *et al.* *J. Clim.* **14**, 2238–2249 (2001).

5. Jones, P. D. *Int. J. Climatol.* **11**, 585–607 (1991).

6. Jones, J. M. & Widmann, M. *J. Clim.* **16**, 3511–3524 (2003).

7. Marshall, G. J. *J. Clim.* **16**, 4134–4143 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Corrigendum

Arrival synchrony in migratory birds

T. G. Gunnarsson, J. A. Gill, T. Sigurbjörnsson, W. J. Sutherland
Nature **431**, 646 (2004).

The line in Fig. 1a shows unity ($x=y$) and is not a regression line, as the legend describes it.

brieF communications arising online

► www.nature.com/bca

Copper oxide superconductors: Sharp-mode coupling in high- T_c superconductors

T. Cuk, Z.-X. Shen, A. D. Gromko, Z. Sun & D. S. Dessau
(doi:10.1038/nature03163)

Reply: J. Hwang, T. Timusk & G. D. Gu
(doi:10.1038/nature03164)

Asteroseismology: Oscillations on the star Procyon

F. Bouchy, A. Maeder, M. Mayor, D. Mégevand, F. Pepe, D. Sosnowska (doi:10.1038/nature03165)

Copper oxide superconductors

Sharp-mode coupling in high- T_c superconductors

Arising from: J. Hwang, T. Timusk & G. D. Gu *Nature* **427**, 714–717 (2004)

In conventional superconductivity, sharp phonon modes (oscillations in the crystal lattice) are exchanged between electrons within a Cooper pair, enabling superconductivity. A critical question in the study of copper oxides with high critical transition temperature (T_c) is whether such sharp modes (which may be more general, including, for example, magnetic oscillations) also play a critical role in the pairing and hence the superconductivity. Hwang *et al.* report evidence that sharp modes (either phononic or magnetic in origin) are not important for superconductivity in these materials¹, but we show here that their conclusions are undermined by the insensitivity of their experiment to a crucial physical effect^{2–7}.

The optics experiment performed by Hwang *et al.* measures a momentum average and is therefore not a sensitive probe when the signal is strongly momentum dependent, as it is for these materials. Existing angle-resolved photoemission (ARPES) data show that in the strongly overdoped regime (with $T_c = 58$ K) there is a prominent 'kink' that is indicative of a peak in the self-energy² (Fig. 1). Figure 1b reveals a clear dispersion kink in the superconducting state near 40 meV (arrow). The strong presence of the

mode signal in this comparably overdoped sample is in contradiction of the central claim of Hwang *et al.*¹.

Figure 1c, d shows that the kink strength (or the peak height of the extracted self-energy, $\text{Re}\Sigma$) from an overdoped sample with T_c of about 71 K is hardly detectable near the node, but is quite strong near the antinode. Hwang *et al.* make no mention of the clear, positive ARPES signal at the antinode, which would otherwise have ruled out their conclusion¹.

T. Cuk*, Z.-X. Shen*, A. D. Gromko†, Z. Sun†, D. S. Dessau†

*Department of Applied Physics, Physics and Stanford Synchrotron Radiation Laboratory, Stanford University, Stanford, California 9430, USA

e-mail: tanjacuk@stanford.edu

†Department of Physics, University of Colorado, Boulder, Colorado 80309, USA

doi:10.1038/nature03163

1. Hwang, J., Timusk, T. & Gu, G. D. *Nature* **427**, 714–717 (2004).
2. Gromko, A. D. *et al. Phys. Rev. B* **68**, 174520 (2003).
3. Kaminski, A. *et al. Phys. Rev. Lett.* **86**, 1070–1073 (2001).
4. Sato, T. *et al. Phys. Rev. Lett.* **91**, 157003 (2003).
5. Abanov, A., Chubukov, A. V., Eschrig, M., Norman, M. R. & Schmalian, J. *Phys. Rev. Lett.* **89**, 177002 (2002).
6. Cuk, T. *et al. Phys. Rev. Lett.* **93**, 117003 (2004).
7. Devereaux, T. P., Cuk, T., Shen, Z.-X. & Nagaosa, N. *Phys. Rev. Lett.* **93**, 117004 (2004).

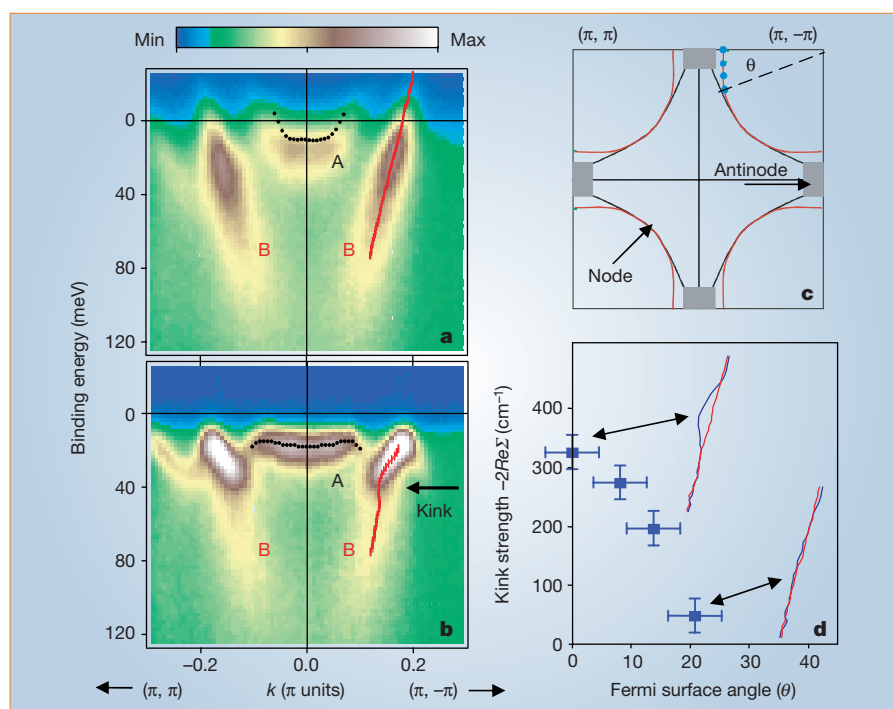


Figure 1 Angle-resolved photoemission (ARPES) data showing a kink in a heavily overdoped $T_c = 58$ K Bi2212 sample (see ref. 2 for details of variables, symbols and coloration). **a**, Normal-state data ($T = 85$ K) from an overdoped sample near the antinodal region (**c**). **b**, Superconducting-state data from the same sample at 10 K, showing the emergence of a dispersion kink in the bilayer split-B band (arrow). **c**, **d**, Momentum dependence of the strength of the temperature-dependent kink (the real part of the self-energy Σ , taking the normal-state curve as reference) from an overdoped $T_c = 71$ K sample (**d**), with locations indicated on the Brillouin zone (**c**). The normal (red) and superconducting (blue) dispersion curves for the extreme locations are shown as well. The ARPES spectra discussed in ref. 1 were taken at 45° (the node, for example). Reprinted with permission from ref. 2; copyright (2003) of the American Physical Society.

Hwang *et al.* reply — Our optical technique has the advantage of being a bulk probe, which is less subject to uncertainties in the doping level and in the quality of the surface than ARPES. It is also capable of higher energy resolution and the overall noise level is lower. The disadvantage is that it gives momentum-averaged properties.

In light of these differences, it came as a surprise to us that our reported optical self-energies¹ were able to track in accurate detail the ARPES self-energies of Johnson *et al.*². Our data indicate that, as a function of doping, not only could both optical and ARPES techniques resolve the sharp mode from the background but also that the sharp-mode intensity decreases uniformly, disappearing completely at a doping level of 0.23. As superconductivity is still strong at this doping level, with a T_c of 55 K, we conclude that the sharp mode is not an important contributor to high-temperature superconductivity.

Cuk *et al.*³ make the points that optical data may be insensitive to strongly momentum-dependent signals because they are momentum-averaged, and also that in their ARPES data⁴ for momenta near the antinodal point ($\pi, 0$), the sharp resonance persists in the highly overdoped region and does not disappear as we claim.

Although the measurements of Johnson *et al.*² were performed at the nodal point, the weakening of the resonance also takes place at the antinodal point, as indicated by other ARPES work³. As shown in Fig. 2, self-energy effects at $(\pi, 0)$ are strongly doping dependent, joining the normal-state background at a doping level of 0.24 — just as they do in our optical results and in the ARPES data of Johnson *et al.*² at the nodal point. All three experiments show the same strong doping dependence.

It is therefore surprising that the work of Gromko *et al.*⁴ fails to confirm these results. These authors do not present doping-dependent plots of the self-energy, but a visual inspection of Fig. 2 of ref. 4 suggests that the self-energy effects are almost doping

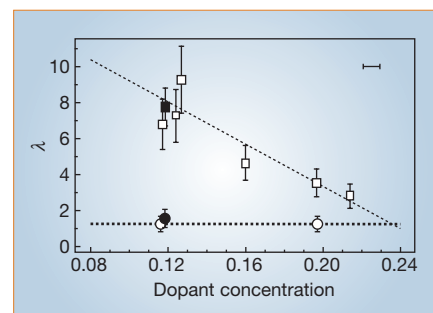


Figure 2 The coupling-strength parameter λ as a function of dopant concentration (see Fig. 3 in ref. 5). Squares, superconducting state; circles, normal state; open symbols, bonding band; filled symbols, antibonding band. Dashed lines are straight-line fits to the data. Horizontal bar, experimental error in the dopant concentration. Reprinted with permission from ref. 5; copyright (2003) of the American Physical Society.

independent — in contrast to the marked doping dependence reported by Kim *et al.*⁵. One reason for the disagreement may be the difficulty in controlling the doping level in the surface layers under the ultra-high-vacuum conditions used in the ARPES experiments.

Jungseok Hwang*, Thomas Timusk*, Genda D. Gu†

*Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L9H 5M3, Canada
e-mail: timusk@mcmaster.ca

†Department of Physics, Brookhaven National Laboratory, Upton, New York 11973-5000, USA
doi:10.1038/nature03164

1. Hwang, J., Timusk, T. & Gu., G. D. *Nature* **427**, 714–717 (2004).
 2. Johnson, P. D. *et al. Phys. Rev. Lett.* **87**, 177007 (2001).
 3. Cuk, T., Shen, Z.-X., Gromko, A. D., Sun, Z. & Dessau, D. S. *Nature* doi:10.1038/nature03163 (2004).
 4. Gromko, A. D. *et al. Phys. Rev. B* **68**, 174520 (2003).
 5. Kim, T. K. *et al. Phys. Rev. Lett.* **91**, 167002 (2003).
- American Physical Society material:** Readers may view, browse, and/or download material for temporary copying purposes only, provided these uses are for noncommercial personal purposes. Except as provided by law, this material may not be further reproduced, distributed, transmitted, modified, adapted, performed, displayed, published, or sold in whole or part, without prior written permission from the publisher.

Asteroseismology

Oscillations on the star Procyon

Arising from: J. M. Matthews *et al. Nature* **430**, 51–53 (2004)

Stars are spheres of hot gas whose interiors transmit acoustic waves very efficiently. Geologists learn about the interior structure of Earth by monitoring how seismic waves propagate through it and, in a similar way, the interior of a star can be probed using the periodic motions on the surface that arise from such waves. Matthews *et al.* claim that the star Procyon does not have acoustic surface oscillations of the strength predicted¹. However, we show here, using ground-based spectroscopy, that Procyon is oscillating, albeit with an amplitude that is only slightly greater than the noise level observed by Matthews *et al.* using spaced-based photometry.

The new spectrograph HARPS² (for High-accuracy Radial-velocity Planet Searcher), which was installed last year on the 3.6-metre telescope of the European Southern Observatory (La Silla, Chile), was optimized for

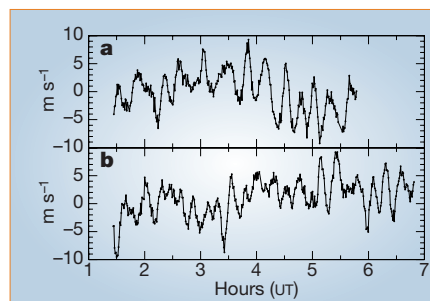


Figure 1 Short sequences of radial-velocity measurements made on Procyon with HARPS spectrograph. **a, b**, Data were collected on **a**, 5 January 2004 and **b**, 6 January 2004. These sequences indicate oscillation modes with periods of around 18 min. UT, universal time.

accuracy in Doppler measurement in order to detect extrasolar planets by means of radial-velocity measurements. During its commissioning, we tested its short-time precision on a sample of bright solar-type stars, including Procyon.

Our measurements on Procyon indicate that there are periodic oscillations of its surface that have a typical period of 18 min (Fig. 1). The apparent amplitude of 4–6 m s^{−1} does not correspond to individual *p*-mode amplitudes, considering that several tens of *p*-modes with similar periods are presumably interfering. Figure 2 presents the Fourier amplitude spectra of the two short time-series obtained on Procyon. No filtering has been applied to the data. Several peaks appear between 0.5 and 1.5 mHz and present the clear signature of acoustic oscillation modes. The correspondence of the main peaks around 1 mHz strongly support the reality of this signature.

Our frequency resolution, which is about 0.55 mHz, does not allow us to resolve individual *p*-modes. The amplitudes of the peaks between 1.0 and 1.5 m s^{−1} probably correspond to two or three times the amplitude of individual modes. The mean white-noise level above 2 mHz is respectively 0.11 m s^{−1} and 0.09 m s^{−1} for the first and second sequences. This result, based on only a few hours of observations, confirms and enforces the previous Doppler ground-based detections^{3–6}.

Why did the Canadian MOST (for Microvariability and Oscillations of Stars) space mission¹ not detect any signatures of *p*-modes on Procyon? The typical amplitude

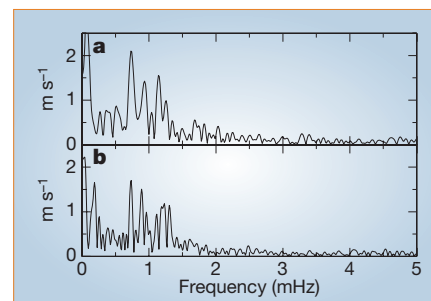


Figure 2 Fourier amplitude spectra of the two short sequences made on Procyon. **a, b**, Signatures of *p*-modes in the frequency range of 0.5–1.5 mHz are evident.

of *p*-modes is about 0.5 m s^{−1} in radial velocity and the relation for converting between velocity and luminosity amplitudes (given by equation (5) of ref. 7) predicts a luminosity amplitude of only 8–10 p.p.m. This is only slightly greater than the noise level of the satellite obtained after 768 hours of observation. These results indicate that the MOST data are dominated by non-stellar noise, as suspected⁸.

However, this conclusion should not overshadow the scientific importance of the Canadian satellite. MOST will lead to breakthroughs on stars with higher oscillation amplitudes, as well as on fast-rotating stars that are not suitable for spectroscopic measurement. The result obtained with HARPS demonstrates the potential of ground-based Doppler measurements for asteroseismology. But for uninterrupted listening to stellar music, a spectrograph like HARPS located in Dome C in Antarctica or in space is needed.

François Bouchy*, André Maeder†, Michel Mayor†, Denis Mégevand†, Francesco Pepe†, Danuta Sosnowska†

*Laboratoire d'Astrophysique de Marseille, Traverse du Siphon, BP8, 13372 Marseille Cedex 12, France
e-mail: francois.bouchy@oamp.fr

†Observatoire de Genève, 1290 Sauverny, Switzerland
doi:10.1038/nature03165

1. Matthews, J. M. *et al. Nature* **430**, 51–53 (2004).
2. Mayor, M. *et al. The Messenger* **114**, 20–24 (2003).
3. Brown, T. M., Gilliland, R. L., Noyes, R. W. & Ramsey, L. W. *Astrophys. J.* **368**, 599–609 (1991).
4. Martić, M. *et al. Astron. Astrophys.* **351**, 993–1002 (1999).
5. Martić, M., Lebrun, J.-C., Appourchaux, T. & Korzennick, S. G. *Astron. Astrophys.* **418**, 295–303 (2004).
6. Eggenberger, P., Carrier, F., Bouchy, F. & Blecha, A. *Astron. Astrophys.* **422**, 247–252 (2004).
7. Kjeldsen, H. & Bedding, T. R. *Astron. Astrophys.* **293**, 87–106 (1995).
8. Christensen-Dalsgaard, J. & Kjeldsen, H. *Nature* **430**, 29–30 (2004).

nature insight

cell division & cancer



The development of cancer can be viewed as an evolutionary process. Cells are constantly subject to mutations in their DNA which are usually detrimental to the cell. But occasionally these changes produce cells that can escape the normal constraints and flourish as pathological tumours.

Cancer cells are selected for their ability to divide when they shouldn't, trigger their own blood supply to support unlimited expansion, and invade the bloodstream and other tissues to form fatal metastases. Changes in the cell-cycle and apoptotic machineries, or in the signalling pathways that control them allow cancer cells to escape the normal control of cell proliferation and cell death.

There is also a growing recognition that changes in the microenvironment of cancer cells can promote their proliferation. Moreover, genomic instability caused by faulty cell division or defective DNA repair may increase the rate of potentially tumorigenic mutations and so contribute to cancer evolution.

Cancer is a complex disease. Enormous heterogeneity in the genetic changes and the context in which they affect cancer development and progression makes it difficult to design effective treatments. Understanding and exploiting these complexities holds promise for more effective therapies in the future. Moreover, the notion that tumour maintenance critically depends on a small subset of cells with stem-cell-like behaviour may mean that a cancer cure ultimately has to target these cells.

Paradoxically, the altered cellular networks of molecular pathways that sustain cancer cell growth and make them resistant to certain therapies may offer new targets for therapy. Critical signalling junctions may exist that are more important for the survival of cancerous than normal cells.

We hope the articles in this Insight capture the excitement and promise of this rapidly advancing field. We are grateful to the authors for their contributions and to the reviewers for their valuable input.

We are pleased to acknowledge the financial support of AstraZeneca in producing this Insight. As always, *Nature* carries sole responsibility for the editorial content and peer review.

Barbara Marte Senior Editor

introduction

- 294 Targeted cancer therapy**
C. Sawyers

review articles

- 298 G1 cell-cycle control and cancer**
J. Massagué

- 307 Intrinsic tumour suppression**

S. W. Lowe, E. Cepero & G. Evan

- 316 Cell-cycle checkpoints and cancer**

M. B. Kastan & J. Bartek

- 324 Tissue repair and stem cell renewal in carcinogenesis**

P. A. Beachy, S. S. Karhadkar & D. M. Berman

- 332 Stromal fibroblasts in cancer initiation and progression**

N. A. Bhowmick, E. G. Neilson & H. L. Moses

progress

- 338 Aneuploidy and cancer**
H. Rajagopalan & C. Lengauer

Editor, *Nature*: Philip Campbell
Insight Publisher: Sarah Greaves
Insight Editor: Lesley Anson
Editorial Assistant: Timothy Gibbs

Production Editor: Lucy Odling-Smee
Art Editor: Martin Harrison
Layouts: Nadia Rooney
Diagrams: Nik Spencer

Sponsorship: Claire Hines
Production: Sue Gray
Marketing: Claire Aspinall

Targeted cancer therapy

Charles Sawyers

Howard Hughes Medical Institute, David Geffen School of Medicine at UCLA, Jonsson Comprehensive Cancer Center, 10833 LeConte Avenue, Los Angeles, California 90095, USA (e-mail: csawyers@mednet.ucla.edu)

Disruption of the normal regulation of cell-cycle progression and division lies at the heart of the events leading to cancer. Complex networks of regulatory factors, the tumour microenvironment and stress signals, such as those resulting from damaged DNA, dictate whether cancer cells proliferate or die. Recent progress in understanding the molecular changes that underlie cancer development offer the prospect of specifically targeting malfunctioning molecules and pathways to achieve more effective and rational cancer therapy.

The reviews in this Insight summarize our current understanding of how oncogene and tumour suppressor gene networks influence the decisions of cancer cells to proliferate or die (Fig. 1). These decisions are further influenced by the tumour microenvironment and stress signals, such as DNA damage. Moreover, recent work suggests that a sub-population of cancer cells with stem-cell-like properties may be critical for triggering tumour development. Together, these studies provide a conceptual framework within which practitioners of experimental cancer therapeutics can consider the design of targeted agents. The term 'targeted therapy' refers to a new generation of cancer drugs designed to interfere with a specific molecular target (typically a protein) that is believed to have a critical role in tumour growth or progression. The identification of appropriate targets is based on a detailed understanding of the molecular changes underlying cancer. This approach contrasts with the conventional, more empirical approach used to develop cytotoxic chemotherapeutics — the mainstay of cancer drug development in past decades. Here, I summarize current progress in targeted therapy, and review the potential targets that are emerging. I focus in particular on kinases, which have so far proved to be a promising class of targets for cancer therapy.

Targeting mutant kinases

The clinical success of the small molecule kinase inhibitor imatinib mesylate (Gleevec) in chronic myeloid leukaemia (CML) and gastrointestinal stromal tumours (GIST) has established a paradigm for the treatment of tumours whose growth is acutely dependent on specific kinase targets (Table 1). CML is driven by the mutant kinase fusion protein Bcr–Abl, which displays constitutive activation of the Abl kinase, whereas GIST is caused by activating point mutations in the c-Kit or platelet derived growth factor receptor (PDGFR)- α kinases. Imatinib effectively blocks the activity of all three kinases and produces dramatic clinical responses in all three situations in a manner that correlates precisely with the presence of these mutations in the tumour¹. In lung cancer, clinical responses to epidermal growth factor receptor (EGFR) inhibitors are associated with point mutations in the EGFR kinase domain^{2,3} (thereby explaining the rather modest 10% response rate in all patients). The clear prediction from this experience is that clinical responses to kinase inhibitors occur in tumours bearing activating mutations that drive tumour progression. Extending this paradigm to larger numbers of cancer patients would require establishing the frequency of kinase mutations in human cancer on a much broader scale — presumably through global gene sequencing efforts analogous to the genome project. Indeed, initial efforts from groups at the

Sanger Centre, the Eli Broad Institute and Johns Hopkins have found previously unsuspected kinase mutations in human tumours^{4–7}.

Given that the fraction of human cancers known to have kinase-domain mutations is currently small, can we realistically expect a substantial percentage of all cancer patients to benefit from kinase inhibitors? There are several reasons for optimism. First, the high frequency of B-Raf kinase mutations in melanoma was completely unrecognized until they were uncovered by a systematic sequencing effort⁵. Hence, many more surprises may follow. Second, serendipitous clinical responses in patients with rare conditions has led to the discovery of previously unrecognized mutant kinases in diseases such as the FIPL1–PDGFR α fusion in hypereosinophilic syndrome⁸. Third, clinical responses are also observed when tumours contain mutations in genes that activate the kinase indirectly. For example, a chromosome translocation causes overproduction of the kinase ligand PDGF in dermatofibrosarcoma protuberans⁹. This last example underscores the fact that global surveys intended to define the frequency of kinase-dependent cancers must also consider the multitude of indirect mechanisms that could lead to constitutive kinase activation.

Caveats for therapies targeting mutant kinases

Assuming that cancer genome surveys reveal that a large fraction of human cancers have kinase-pathway abnormalities amenable to pharmacological blockade, additional complexities may temper expectations for Gleevec-like results in all cancers. One consideration is whether a kinase-pathway mutation occurs early or late in the life history of the tumour, as this may affect the degree to which tumour growth is dependent on these changes. It is generally agreed that the Bcr–Abl mutation in CML serves as the initiating event, raising the question of whether the dramatic clinical responses seen in this disease are, in part, attributable to attacks on the earliest oncogenic lesion. In contrast, there is some evidence to suggest that another kinase abnormality involving the Flt3 receptor in acute myeloid leukaemia occurs late in disease progression. Phase I clinical data with at least three different Flt3 inhibitors has provided clear evidence of tumour response, but the magnitude of the response appears less impressive than that induced by Gleevec in late-stage CML (refs 10–12). Of course, a large number of other variables, such as efficacy of target inhibition, could explain this difference, but the chronology of where the targeted lesion abnormality occurs in the scheme of tumour progression could have a role in clinical outcome.

An additional consideration is disease relapse due to drug resistance. Perhaps the best understanding of this problem at a molecular level comes from studies of Gleevec resistance in CML patients. Relapse is caused by the expansion of tumour

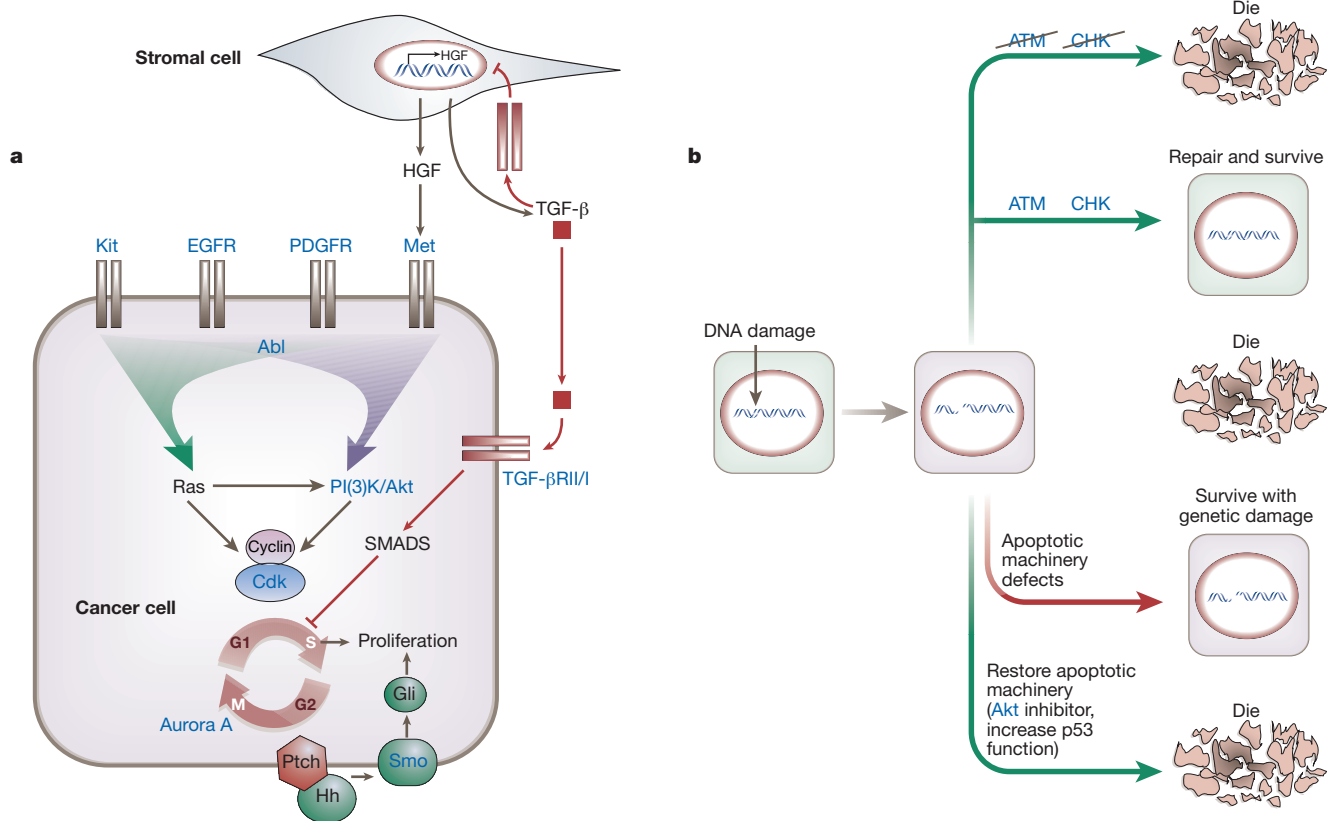


Figure 1 Cancer pathways and targeted therapy. **a**, Multiple signalling pathways — upregulated in cancer cells owing to specific alterations in oncogenes or tumour suppressors — stimulate tumour-cell proliferation, often by promoting G1–S cell-cycle progression (see reviews in this issue by Massagué, page 298; and by Beachy *et al.*, page 324). Signals from the tumour microenvironment, including stromal fibroblasts (see review in this issue by Bhowmick *et al.*, page 332), can positively or negatively shape cancer-cell proliferation. The inhibition of growth-promoting pathways by therapy tailored to the specific genetic alterations found in cancer offers a new therapeutic approach: an example is the recent approval of drugs targeting the Abl and EGFR kinases. These and other potential drug targets are shown in blue. CIN and other mitotic defects leading to aneuploidy (see progress article in this issue by Rajagopalan and Lengauer, page 338) may cause the accumulation of genetic

defects endowing cancer cells with a selective growth advantage and providing another possible route for treatment. One potential therapeutic target is the Aurora A kinase, which is implicated in mitotic progression and CIN. **b**, Classical chemotherapy and radiotherapy eliminates cancer cells by inducing DNA damage and subsequent apoptosis. DNA-damage-response pathways (which employ the ATM and CHK1/2 kinases) promote repair and survival (see review in this issue by Kastan and Bartek, page 316). Defects in the apoptotic machinery can allow cancer cells to survive DNA damage, which may lead to the acquisition of further mutations (see review in this issue by Lowe *et al.*, page 307). Inhibition of DNA-damage-response pathways or restoration of defective apoptosis pathways may render cancer cells more susceptible to DNA-damaging agents and provide potential avenues for more efficient and tumour-specific future therapies in the future.

subclones in the face of continued therapy; these subclones contain single amino-acid mutations in the Bcr–Abl kinase domain that prevent enzyme inhibition by Gleevec^{13–15}. Recent preclinical studies have identified second generation dual Src/Abl kinase inhibitors that retain activity against nearly all the Gleevec-resistant mutants¹⁶. These compounds are now in early clinical testing, but the expectation is that future therapy will rely on cocktails of inhibitors to prevent the emergence of resistant subclones.

Extending lesion-specific therapy to pathways

The above discussion exemplifies recent clinical success in cancer therapy by directly targeting the oncogenic lesions responsible for tumour initiation and progression. But can this approach be exploited more broadly, by inhibiting pathway components that are not themselves mutated, even though the oncogenic lesion occurs in a pathway regulator? The example of dermatofibrosarcoma protuberans discussed above represents one proof-of-concept example in a rare disease, but can we realistically expect a more generalized impact for this approach? Several of the reviews in this issue provide reason for optimism.

Patients with Gorlin's syndrome have an inherited predisposition to develop cancers because of germline mutation in the Patched

Table 1 Targeted agents and their current status in clinical testing

Drug	Target	Disease	Clinical trial status
Imatinib (Gleevec)	Abl Kit PDGFR	CML GIST HES CMML DFSP	Approved
Gefitinib (Iressa)	EGFR	Lung cancer	Approved
Bevacizumab (Avastin)	VEGF ligand	Colon cancer	Approved
CCI-779 RAD-001	mTOR	Various cancers	Phase I, II, III
BMS-354825	Abl KIT	CML GIST	Phase I
PKC-412 MLN-518 CEP-701	FLT3	AML	Phase I/II
BAY 43-9006	VEGFR RAF	Kidney cancer Melanoma	Phase I/II
SU-011248	VEGFR	Kidney cancer	Phase I/II

AML, acute myeloid leukaemia; HES, hypereosinophilic syndrome; CMML, chronic myelomonocytic leukaemia; DFSP, dermatofibrosarcoma protuberans.

(Ptc) gene, a cell surface receptor that functions in the Hedgehog (Hh) pathway (see Fig. 1 in review by Beachy *et al.*, page 324). The natural product cyclopamine interferes with the downstream Hh-pathway protein Smoothened (Smo) and impairs the growth of such tumours in model systems. Remarkably, the usefulness of this approach may not be limited to the small fraction of tumours with Ptc mutation. In many epithelial tumours, such as those arising in the pancreas and prostate, Hh-pathway ligands are produced, resulting in constitutive pathway activation and, most importantly, cyclopamine sensitivity. The molecular lesion(s) leading to ligand-dependent activation in these tumours have not been defined. In at least some cases, there seems to be a tumour-specific molecular event that alters the availability of Smo for downstream signalling. Regardless of the mechanism underlying pathway activation, these findings have already motivated the pharmaceutical industry to search for Hh-pathway inhibitors, one of which has recently been shown to be effective in a Hh-pathway-dependent mouse model of medulloblastoma¹⁷.

Since the elucidation of the role of cyclin-dependent kinases (CDKs) in cell-cycle regulation, these proteins have been extensively explored as potential drug targets. Highly selective and potent CDK inhibitors exist, but the overriding question is how we define those tumours in which the inhibition of CDK activity provides a favourable therapeutic index. If we use the pathway-mutation paradigm to address this question, tumours with molecular lesions in the primary cell-cycle machinery should be susceptible to CDK inhibition. At a minimum, this might include rare familial melanomas with mutations in CDKs (ref. 18), mantle cell lymphomas with translocations leading to increased cyclin D1 expression¹⁹, and tumours with loss-of-function mutations in p16 or retinoblastoma protein (Rb). However, the list could be much larger as a number of the signalling pathways more commonly mutated in human cancers (Myc, Ras, transforming growth factor beta (TGF- β) through receptor mutations or Smad 4 loss), impinge directly on cell-cycle regulation (see review in this issue by Massagué, page 298). The challenge in clinical evaluation of these inhibitors is to design trials that enroll patients for whom there is a detailed knowledge of the molecular phenotype of the tumour, so that appropriate correlations with clinical responses can be made.

A third example is to use inhibitors of kinases in the PI(3)K/Akt/mTOR pathway to treat tumours with mutations in the tumour suppressor gene *PTEN* — the negative principal regulator of this pathway (see reviews in this issue by Massagué, page 298, and by Lowe *et al.*, page 307). Proof of concept for this approach has been demonstrated in numerous murine models using rapamycin analogues that block mTOR activity²⁰. Because mTOR receives signalling inputs from several signalling pathways, tumours with a number of distinct molecular lesions could be sensitive to treatment. As with CDK inhibitors, the current challenge is to correlate clinical activity with molecular phenotype.

Targeting cancer stem cells

Even in the seemingly near-perfect world of Abl-kinase-inhibitor treatment of CML, it is becoming increasingly clear that we cannot ignore the concept of cancer stem cells. As discussed by Beachy and colleagues (in this issue, page 324), there is growing evidence that some, if not all, tumours derive from a small number of stem-cell-like cells that either retain or acquire the capacity for self-renewal. Presumably, targeted therapies must eliminate tumour stem cells to prevent a later relapse. The clinical experience with imatinib in CML provides an opportunity to consider this concept in patients. Despite the fact that imatinib reliably reduces the tumour burden in CML by three to four orders of magnitude, most patients continue to have residual disease. This is detected by quantitative polymerase chain reaction (PCR) for the Bcr-Abl fusion breakpoint²¹. The risk of relapse in these patients remains low with three years of clinical follow up, but recent studies suggest that these residual CML cells

reside in the stem-cell-like CD34+ population, and may contain imatinib-resistance mutations^{22–24}. Persistence in this self-renewing pool raises obvious concerns about the eventual emergence of resistant subclones. The new dual Src/Abl inhibitors discussed above, which are more than 100-fold more potent than imatinib, could theoretically represent a solution, provided that the CML stem cell requires Bcr-Abl for survival.

On the basis of our current understanding of cancer stem cells, can we envision a strategy to eliminate them? If these cells have unique patterns of cell-surface antigen expression, monoclonal antibodies might be designed to target them specifically (assuming these antigens would not be shared by normal stem cells from the tissue of origin). Perhaps more promising is the potential to target specific signalling pathways required for stem-cell function. The most likely suspects, on the basis of current thinking, are the Hh and Wnt pathways, both of which have oncogenic potential based on known mutations in pathway components found in several human tumours (see review in this issue by Beachy *et al.*, page 324). Furthermore, pharmacological blockade of these pathways (like that demonstrated with cyclopamine) seems feasible.

How might such inhibitors perform in clinical trials? If the effects are relatively stem-cell specific, clinical responses may be slow to manifest. Consider, by comparison, a drug that affects the large mass of more differentiated tumour cells (imatinib), which shows clinical responses in days to weeks. The effects of a stem-cell drug may take longer to become clinically evident (possibly months or years), especially if the differentiated tumour cells that are not affected by the therapy have a long lifespan as is the case with some epithelial tissues. An additional consideration is safety; extended monitoring for delayed toxicity due to loss of normal stem-cell function in the relevant organ may be needed.

Targeting the microenvironment

Much attention over the years has been devoted to the notion that the tumour blood supply can be targeted with antiangiogenic agents. This has culminated in the recent approval of a monoclonal antibody directed against the vascular endothelial growth factor (VEGF) ligand (which is essential for endothelial cell proliferation) for the treatment of colon cancer, when used in conjunction with chemotherapy²⁵. Part of the attraction of this approach is the near universality of its potential application, as essentially all cancers require a blood supply for their continued growth and spread. In addition, there is the notion that therapy directed against the supporting host tissue rather than the tumour itself will be less prone to resistance because the genetic plasticity of the cancer is not reflected in the stroma. Curiously, the clinical activity of the VEGF antibody in colon cancer was not anticipated and does not obviously correlate with tumour-associated angiogenesis patterns. Notably, the VEGF antibody and small molecule inhibitors targeting the VEGF tyrosine kinase receptor have both shown impressive single-agent activity in renal cancer^{26–28}. These tumours are highly vascular owing to the deletion of the von Hippel-Lindau tumour suppressor gene — the primary molecular lesion in these cancers. This leads to upregulation of the HIF transcription factors and constitutive expression of VEGF in tumour cells²⁹. The growth of these tumours is driven by HIF, in much the same way as activating kinase mutations drive the growth of some of the cancers discussed above, so the anti-tumour properties of VEGF-pathway drugs may not occur solely through effects on the stroma because these tumours often express VEGF receptors.

Recent advances in our understanding of tumour-stroma interaction reveal a much more complex interplay that extends well beyond the simple notion of vascularity. Tumour stroma includes stromal fibroblasts and a number of different inflammatory cells that can clearly modulate tumour growth. Among the best-studied stromal factors is TGF- β , which can exert myriad effects that influence tumour growth in a positive or negative fashion (see also review in this issue by Bhowmick *et al.*, page 332). Although TGF- β has

immunosuppressive properties that may hamper host immune surveillance, it also exerts direct anti-proliferative effects on epithelial cells by engaging the TGF- β receptor (TGF β R)/SMAD signalling pathway. Remarkably, the tumour-suppressive effect of TGF- β is critical in stromal cells as well as in adjacent epithelium, as was recently demonstrated by the finding that epithelial malignancy can develop in certain organs when the TGF- β receptor is deleted only in stromal fibroblasts³⁰. This result, together with earlier studies that demonstrate pro-oncogenic characteristics of cancer-associated fibroblasts isolated from human prostate tumours³¹, clearly establishes that stroma should not merely be considered as supportive to the cell-autonomous growth of tumour cells. Instead, the stroma can exert profound effects on the initiation and progression of epithelial malignancies. Elucidation of the molecular circuitry of this crosstalk could profoundly influence our thinking about targeted cancer therapy, and may provide new prevention strategies.

Combining classical chemotherapy with targeted therapy

Although the above discussion offers many exciting new targets for treating and/or preventing cancer, classical chemotherapy and radiotherapy approaches remain the mainstay of cancer treatment for tumours that cannot be cured solely by surgical excision. Because of the recent success in defining the biochemical details of cells' responses to DNA-damaging agents, these conventional cancer therapies may be combined with targeted agents that disrupt the DNA-repair response; hopefully a more catastrophic cell kill in tumours will result. Proof of concept comes from the well-known hypersensitivity of ataxia-telangiectasia patients to ionizing radiation and chemotherapy. The molecular basis for this exaggerated response is deficiency in DNA-damage-induced cell-cycle arrest owing to mutation in ATM kinase in these patients (see review in this issue by Kastan and Bartek, page 316). This compromises the time needed to repair DNA lesions that are induced by chemo- or radiotherapy. It stands to reason that pharmacological inhibition of ATM or downstream CHK kinases with specific inhibitors might similarly impair the DNA-damage response to conventional cancer therapy, and provoke an even greater apoptotic response. The challenge for pursuing this approach will be to ensure an adequate therapeutic index, such that the nearly universal toxicities of chemotherapeutic agents on normal haematopoietic and gastrointestinal epithelial cells are not similarly enhanced. One scenario might be to use ATM or CHK kinase inhibitors in combination with focal radiotherapy, such that DNA damage is restricted only to cells in the radiation-therapy field.

The above discussion addresses the goal of maximizing tumour cell kill with conventional cancer therapy by poisoning the cell's ability to repair the damage induced by these agents. An alternative strategy that may achieve a similar goal is to define precisely why some tumours fail to respond to chemotherapy in the first place, and then to interfere with these resistance pathways so that cytotoxics can provoke tumour shrinkage. A vast amount of work on this question suggests that most cancers acquire defects in apoptosis pathways (see review in this issue by Lowe *et al.*, page 307), such that tumour cells fail to die despite the presence of strong pro-apoptotic signals induced by chemotherapy. The cataloguing of human tumours for such defects indicates myriad mechanisms, but there are several nodal points in the pathway that might be amenable to pharmacological intervention. These involve proteins such as Bcl2, p53 or Akt kinase. In each case, small molecule strategies have been reported in model systems that either block (Bcl2, Akt) or promote (p53) the activity of the target protein. These could potentially restore the cell-death response in the presence of DNA damage. Clinical evaluation of any of these concepts, however, has yet to be initiated.

Concluding remarks

The reviews in this Insight provide much food for thought. The era of molecular targeted cancer therapy has clearly arrived, but patients and practitioners are yearning for this approach to have a broader

impact. The successes of the past few years illustrate the power of the approach and should reinforce the need to continue basic studies of the molecular underpinnings of human cancers. The failure to see clinical responses with some targeted agents teaches critical lessons as well. Of utmost importance is the need to define the relevant patient population for clinical trials and therapy through molecular characterization of the tumour. Overcoming this barrier will require the development and widespread adoption of appropriate molecular diagnostic assays. Only then will it be possible to realize the broader potential of targeted cancer therapy. □

doi:10.1038/nature03095

1. Sawyers, C. L. Opportunities and challenges in the development of kinase inhibitor therapy for cancer. *Genes Dev.* **17**, 2998–3010 (2003).
2. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
3. Sordella, R., Bell, D. W., Haber, D. A. & Settleman, J. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science* **305**, 1163–1167 (2004).
4. Stephens, P. *et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431**, 525–526 (2004).
5. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
6. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinase in colorectal cancers. *Science* **300**, 949–950 (2003).
7. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304**, 554 (2004).
8. Cools, J. *et al.* A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. *N. Engl. J. Med.* **348**, 1201–1214 (2003).
9. Rubin, B. P. *et al.* Molecular targeting of platelet-derived growth factor B by imatinib mesylate in a patient with metastatic dermatofibrosarcoma protuberans. *J. Clin. Oncol.* **20**, 3586–3591 (2002).
10. De Angelo, D. J. *et al.* Phase I clinical results with MLN518, a novel FLT3 antagonist: tolerability, pharmacokinetics, and pharmacodynamics. *Blood* **102** (65a), Abstr. 219 (2003).
11. Stone, R. M. *et al.* Oral PKC 412 has activity in patients (pts) with mutant FLT3 acute myeloid leukemia (AML): a phase II trial. *Proc. Am. Soc. Clin. Oncol. Abstr.* **22**, 563 (2003).
12. Smith, B. D. *et al.* Single agent CEP-701, a novel FLT3 inhibitor, shows initial response in patients with refractory acute myeloid leukemia. *Proc. Am. Soc. Clin. Oncol. Abstr.* **22**, 194 (2003).
13. Gorre, M. E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876–880 (2001).
14. Shah, N. P. *et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2**, 117–125 (2002).
15. Branford, S. *et al.* High frequency of point mutations clustered within the adenosine triphosphate-binding region of BCR/ABL in patients with chronic myeloid leukemia or Ph-positive acute lymphoblastic leukemia who develop imatinib (STI571) resistance. *Blood* **99**, 3472–3475 (2002).
16. Shah, N. P. *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399–401 (2004).
17. Romer, J. T. *et al.* Suppression of the Shh pathway using a small molecule inhibitor eliminates medulloblastoma in Ptc1(+/-)p53(-/-) mice. *Cancer Cell* **6**, 229–240 (2004).
18. Wolfel, T. *et al.* A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. *Science* **269**, 1281–1284 (1995).
19. Williams, M. E., Swerdlow, S. H. & Meeker, T. C. Chromosome t(11;14)(q13;q32) breakpoints in centrocytic lymphoma are highly localized at the bcl-1 major translocation cluster. *Leukemia* **7**, 1437–1440 (1993).
20. Sawyers, C. L. Will mTOR inhibitors make it as cancer drugs? *Cancer Cell* **4**, 343–348 (2003).
21. Hughes, T. P. *et al.* Frequency of major molecular responses to imatinib or interferon alpha plus cytarabine in newly diagnosed chronic myeloid leukemia. *N. Engl. J. Med.* **349**, 1423–1432 (2003).
22. Graham, S. M. *et al.* Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 *in vitro*. *Blood* **99**, 319–325 (2002).
23. Bhatia, R. *et al.* Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood* **101**, 4701–4707 (2003).
24. Chu, S. *et al.* Detection of BCR-ABL kinase mutations in CD34+ cells from chronic myelogenous leukemia patients in complete cytogenetic remission on imatinib mesylate treatment. *Blood* published online 2 September 2004 (doi: 10.1182/blood-2004-03-1114).
25. Hurwitz, H. *et al.* Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *N. Engl. J. Med.* **350**, 2335–2342 (2004).
26. Yang, J. C. *et al.* A randomized trial of bevacizumab, an anti-vascular endothelial growth factor antibody, for metastatic renal cancer. *N. Engl. J. Med.* **349**, 427–434 (2003).
27. Ratain, A. *et al.* ASCO annual meeting proceedings (post-meeting edn). *J. Clin. Oncol.* **22**, (no. 14 suppl.) 4501 (2004).
28. Motzer, R. *et al.* ASCO annual meeting proceedings (post-meeting edn). *J. Clin. Oncol.* **22**, (no. 14 suppl.) 4500 (2004).
29. Kaelin, W. G., Jr. Molecular basis of the VHL hereditary cancer syndrome. *Nature Rev. Cancer* **2**, 673–682 (2002).
30. Bhowmick, N. A. *et al.* TGF-beta signaling in fibroblasts modulates the oncogenic potential of adjacent epithelia. *Science* **303**, 848–851 (2004).
31. Olumi, A. F. *et al.* Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–5011 (1999).

Competing interests statement The author declares that he has no competing financial interests.

G1 cell-cycle control and cancer

Joan Massagué

Cancer Biology and Genetics Program, and Howard Hughes Medical Institute, Box 116, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York 10021, USA (e-mail: j-massague@ski.mskcc.org)

Before replicating DNA during their reproductive cycle, our cells enter a phase called G1 during which they interpret a flood of signals that influence cell division and cell fate. Mistakes in this process lead to cancer. An increasingly complex and coherent view of G1 signalling networks, which coordinate cell growth, proliferation, stress management and survival, is helping to define the roots of malignancies and shows promise for the development of better cancer therapies.

The past three decades have seen an unrelenting quest to understand the reproduction of normal and cancerous cells. Cell reproduction entails replication of the DNA followed by division of the nucleus and partitioning of the cytoplasm to yield two daughter cells. This sequential routine is known as the 'cell cycle' and, as a problem that has fascinated biologists since the mid-nineteenth century, we are only just beginning to understand how it works.

Cells in early embryos can proceed through continuous cycles of DNA replication and nuclear division at astonishing speed. DNA replication starts as soon as mitosis ends and a full cycle of cell division is completed in a mere half hour. But as embryogenesis unfolds and the demands of cell life in a complex environment set in, a bureaucracy arises. A gap called G1 phase is incorporated between nuclear division (M phase) and DNA synthesis (S phase); another gap called G2 phase occurs between S and M (Fig. 1b). These gaps allow for the repair of DNA damage and replication errors. But above all, G1 is a period when many signals intervene to influence cell division and the deployment of a cell's developmental programme. Diverse metabolic, stress and environmental cues are integrated and interpreted during this period. On the basis of these inputs, the cell decides whether to enter S phase or pause. Moreover, in multicellular organisms the behaviour of a cell must obey dictums from its neighbours. To this end, during G1 the cell makes further decisions regarding whether to self-renew, differentiate or die.

The proper interpretation and execution of these inputs is a delicate business and, not surprisingly, mistakes lead to cancer. When mutations allow a cell to remain in a proliferative state and avoid terminal differentiation and death, that cell is poised to escalate in its degenerate behaviour. It can acquire additional mutations for invasion of surrounding tissues and metastatic re-creation of the tumour at distant organs. Many oncogenes and tumour suppressor genes, as well as the therapies that target them, can be linked to faulty G1 control.

How then is the decision to initiate DNA replication controlled in our cells at the molecular level? Ongoing work is providing an ever more complex but coherent answer to this question. Here, I discuss examples of signal transduction pathways that influence G1 progression, I describe how these processes affect cancer formation, and finally, I consider current ideas on potential therapeutic interventions. The recent progress in these areas is so extensive that this review can only distill key principles with little room for personal opinion or comprehensive detail.

A G1 engine of cyclins and kinases

Compared to DNA replication and mitosis, which follow canonical steps that vary little from cell to cell, the steps

controlling entry and progression through G1 are largely dependent on cell type and context. A stem cell that is constantly replenishing the intestinal lining, a lymphocyte suddenly stimulated by antigen, or an angioblast responding to vascular injury, all proceed through G1 phase under different circumstances, different signals, different developmental programmes and with different risks of malignant transformation. Ultimately, however, to enter S phase all cells must fulfill the same essential requirement: they must activate cyclin-dependent kinases (CDKs).

CDKs are protein kinases that require binding to a cyclin subunit to become catalytically competent^{1,2}. Different members of the CDK family, in association with different cyclins, turn key switches throughout the cell cycle; other family members regulate transcription, differentiation, nutrient uptake and other functions. Cyclin-CDK complexes are regulated by phosphorylation and protein interaction events that tightly control the timing and extent of CDK activation. The prototypic CDK, Cdk1, associates with cyclins A and B, and acts at the G2/M interface (Fig. 1). The progressive accumulation of A and B cyclins during the cell cycle and their abrupt degradation at the onset of anaphase, mediates entry and exit from mitosis, respectively. The drop in Cdk1 activity at the end of M phase allows DNA chromosomal sites known as replication origins to be loaded with a pre-replicative complex (PRC) (refs 3, 4; Fig. 2). This complex contains ORC (origin of replication complex), the kinase Cdc6/18 and Ctd1 (Cdc10-dependent transcript 1), and loads MCM (mini-chromosome maintenance) proteins onto the DNA, licensing these sites for the initiation of replication.

G1 CDKs trigger DNA replication. In higher eukaryotes the G1 CDKs include Cdk2, which combines with E-type cyclins (E1, E2) and cyclin A (refs 1, 2). On Cdk2 activation, PRCs recruit DNA helicases, primases and polymerases, causing unwinding of the double helix and DNA replication^{3,4} (Fig. 2). Cdk activity is essential for the unwinding step, and several components of the PRC become phosphorylated in the process. The newly replicated origins cannot reassemble new PRCs until CDK activity once again drops at the end of mitosis. Mitosis in turn will not proceed until DNA replication is completed. Together, these events ensure that DNA will be replicated once and only once, per cell cycle^{3,4}. The identities of the CDK substrates that directly trigger DNA replication remain unknown and a stinging reminder of how much we still do not know about how cell reproduction works.

The scheme summarized here is well supported by experimental evidence, but it cannot be taken too rigidly. For example, E-type cyclins are largely dispensable for mouse development (although cells lacking cyclin E have problems loading MCM onto DNA)⁵. Even more strikingly,

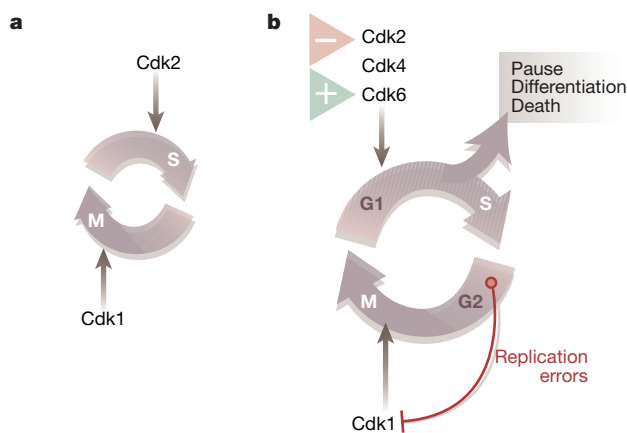


Figure 1 Simple and complex cell cycles **a**, The essential cell cycle such as that which occurs in early embryos comprises nuclear (and cell) division (M phase) under the command of Cdk1, after Cdk2-directed replication of the DNA (S phase). **b**, Later in development and in adult tissues, the cell cycle includes a gap period (G1 phase) during which the activity of various CDKs and other required components is controlled by diverse positive (growth, survival and mitogenic) and negative (apoptotic and cytostatic; genotoxic, metabolic, oncogenic and oxidative stress) signals. Some of these signals come from neighbouring cells or the circulation. Others reflect the metabolic status of the cell, DNA damage caused by genotoxic agents, physical and chemical stresses, or potentially oncogenic stimuli. Another gap period (G2 phase) is devoted to mending replication errors and ensuring that all is in order to proceed with mitosis. Oncogenic transformation is largely the result of malfunctions in these G1 and G2 mechanisms.

mice lacking Cdk2 are viable⁶. It has been argued that the subcellular location and timing of activation of a CDK may be as important as its catalytic specificity⁷. So, compensatory effects by other CDKs might explain these results.

Orchestrating G1 with CDK inhibitors

Cyclin E levels are constantly high in the cells of early embryos, allowing Cdk2 to initiate S phase as soon as M phase is over². In most other cells, however, various mechanisms enforce the existence of G1 phase by keeping Cdk2 inactive until mitogenic signals intervene. One of these mechanisms is based on limiting the supply of cyclin E. Cyclin E expression is dependent on E2F transcription factors^{7,8} (Fig. 2). In mitotically resting cells, and in cells that have just emerged from M phase, E2F factors are bound to the retinoblastoma protein (Rb) or its family members, p107 and p130 (ref. 9). Rb binding turns E2Fs into repressors (in the case of E2F4 and E2F5) or inactive transactivators (in the case of E2F1, -2 and -3)⁸.

Mitogenic stimuli change this state of affairs by increasing the amount of D-type cyclins, which combine with Cdk4 and Cdk6 to phosphorylate and inactivate Rb (refs 1, 2, 10; Fig. 2). Phosphorylation dissociates Rb from E2F, allowing E2F-dependent transcription. Along with cyclin E, the E2Fs activate transcription of a large set of components that support DNA replication (ORCs, MCMs, DNA polymerase α) and subsequent events (cyclin B, Cdk1 and various DNA quality-control components)^{7,8}. Rb seems to be the only essential substrate of cyclin D–Cdk4/6, as cells lacking Rb no longer require cyclin D for proliferation¹¹. Rb can also be phosphorylated by cyclin E–Cdk2, creating a positive feedback loop that helps precipitate S-phase entry once enough Cdk2 has been activated. However, it would be a gross oversimplification to imply that Cdk4/6, Rb, E2F1–3 and Cdk2 function in a strictly linear pathway. Indeed, when ‘knocked-in’ into the cyclin D1 locus, and so placed under the direct control of

mitogenic signals, cyclin E rescues many of the phenotypes observed in cyclin D1 deficient mice¹². Mice that are defective in all three cyclin D genes die at mid/late gestation of severe anaemia, but not without having normally developed many organs and tissues¹³. Thus, haematopoietic cells may be cyclin-D-dependent but many other cell types are ‘cyclin-D-independent’, perhaps indicating that Cdk2 may be sufficient for sensing mitogenic stimuli and mediating proliferation in these cells¹³. A similar phenotype is observed in mice lacking Cdk4 and Cdk6, in which Cdk2 with its usual partner cyclins or with D-type cyclins may be partially compensating for the absence of Cdk4/6 (ref. 14). These observations add to the growing suspicion that, if necessary, different mammalian CDKs may take each other's place in a cell-type-specific manner, or under abnormal circumstances.

Another mechanism that prevents premature entry into S phase, and ties the G1/S transition to regulatory inputs, relies on inhibitory proteins that latch onto cyclin–CDK complexes and disrupt their catalytic centre¹⁵. One of these inhibitors in particular, p27Kip1, functions as an integral brake of the cell cycle (Fig. 2). Others, such as p15Ink4b, p16Ink4a, p21Cip1/WAF1 and p57Kip2 are mediators of cytostatic signals (see section ‘Cytostatic signalling by means of the SMAD node’ below). p27 silences cyclin–Cdk2, which may be present

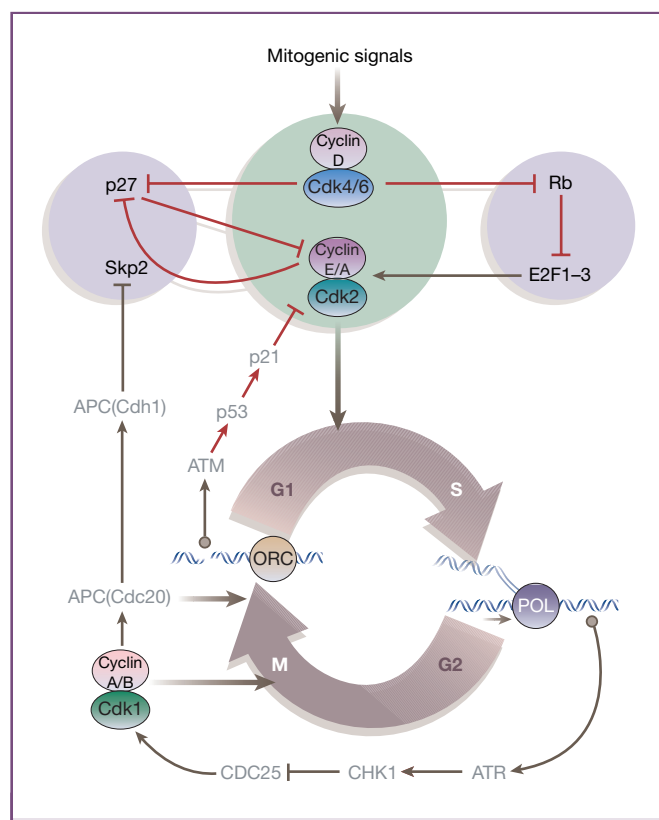


Figure 2 A CDK engine for G1 to S transition and its built-in regulators. At the end of mitosis, PRC bound to DNA (ORC) is ready to initiate replication on the command of Cdk2. A limited supply of cyclins E and A, and the presence of Cdk2 inhibitor p27Kip1, postpone Cdk2 activation. Mitogenic signals acting by means of Cdk4/6 neutralize p27Kip1 and induce E2F-dependent transcription of cyclins and other components, resulting in Cdk2 activation and entry into S phase. Several built-in processes enforce orderly progression of the cell cycle. APC(Cdc20) complexes destroy mitotic cyclins, trigger sister chromatid separation and promote exit from mitosis (M phase). Until the concentration of Cdk1 drops at the end of mitosis, Cdk1 activity prevents ORC reactivation. An APC(Cdh1) complex formed during M phase prevents premature Skp2-dependent destruction of p27Kip1. Damaged DNA activates the ATM–p53 pathway, which prevents CDK activation by means of the inhibitor p21 Cip1/WAF1. Unreplicated DNA activates the ATR–CHK1 pathway, which prevents CDK1 activation and premature mitosis.

in quiescent cells or in early G1. Mitogenic stimuli liberate Cdk2 from inhibition by suppressing the transcription, translation, stability or nuclear localization of p27, or by inducing p27 sequestration by cyclin-D–Cdk4/6 complexes¹⁶. Cdk2 activation is completed by the action of CAK (CDK-activating kinase) and the removal of inhibitory phosphorylation by the phosphatase CDC25. Once the balance tips in favour of Cdk2 activation, Cdk2 bites back by phosphorylating p27 and marking it for polyubiquitination and destruction¹⁷ (Fig. 2).

Indeed, p27 polyubiquitination provides another way to enforce the orderly progression of the cell cycle¹⁸. Polyubiquitination is mediated by a SCF (Skp1/Cul1/F-box protein)-ubiquitin-ligase complex that captures phosphorylated p27 by means of the F-box protein Skp2 and its cofactor Csk1 (ref. 19). The stabilities of Skp2 and Csk1 in turn are controlled by APC/C (anaphase-promoting complex/cyclosome), the other major cell cycle ubiquitin ligase^{19,20}. One form of APC, APC(Cdc20), mediates chromatid separation during M phase and exit from mitosis (by ubiquitinating securin and mitotic cyclins respectively)^{18,21}. A distinct form of APC, containing Cdh1 instead of Cdc20 as the APC activator, is active well into G1 and targets Skp2 for destruction. p27 levels remain high until the activity of APC (Cdh1) subsides later in G1 (Fig. 2). Thus, APC(Cdh1) helps to maintain the gap between M phase and the G1/S transition. As APC (Cdh1) starts running out of substrates in G1, it turns on itself, causing the ubiquitination and degradation of its own ubiquitin-conjugating enzyme subunit²². APC may also be inhibited at the G1/S transition by binding of the inhibitor protein Emi1 (ref. 23). With APC now silenced, S phase can start.

Given their role in orchestrating the G1/S transition, it is not surprising that several components of this network are involved in cancer. Cyclin D1 overexpression, with or without gene amplification, occurs in 50% of breast cancers. In the mouse, cyclin D1 is required for mammary tumour formation by the Ras pathway²⁴ — a potent mediator of growth and survival signals (see next section). The first tumour suppressor gene to be identified encodes Rb (ref. 11). Inherited mutations in one *Rb* allele cause retinoblastoma on somatic inactivation of the other allele. Somatic inactivation of both alleles occurs in various types of cancer. Furthermore, low levels of p27Kip1 correlate with poor prognosis in many types of carcinoma as well as in brain tumours and lymphomas. Increased expression of Skp2 occurs in these tumours¹⁶. In the mouse, p27Kip1 is haplo-insufficient for the suppression of prostate hyperplasia and various malignancies¹⁶. All these alterations may coerce cells into entering S phase instead of differentiating or dying: the resulting expansion of the pool of proliferative cells increases the chances of the cells acquiring additional oncogenic mutations.

Ras and PI(3)K networks power the G1 engine

Work over the past decade has delineated externally activated signalling pathways that control inhibitor and cyclin concentrations, which in turn empower CDKs to drive the G1/S transition. The Ras and PI(3)K (phosphatidylinositol-3-OH kinase) pathways stand out in this respect. Scores of mitogenic factors, acting through many different receptor tyrosine kinases or G-protein-coupled receptors, activate the Ras and PI(3)K pathways to stimulate cell proliferation, growth and survival.

Receptor tyrosine kinases, such as the prototypic epidermal growth factor receptor (EGFR), initiate signalling by creating recruitment centres for effector proteins at the membrane^{25,26} (Fig. 3a). Phosphorylated tyrosine residues on receptor cytoplasmic domains or on surrogate proteins like IRS (insulin-like receptor substrate) create docking sites that are recognized by phosphorylation-binding domains — in adaptor proteins. Adaptors then recruit effectors. The three Ras-family members, H-Ras, K-Ras and N-Ras, are activated in this fashion by means of the adaptors Grb2 and Shc (refs 27, 28). On recruitment to the membrane, Ras proteins undergo cycles of GTP loading by the guanine-nucleotide-exchange

factor SOS, and GTP hydrolysis by the GTPase-activating proteins (GAPs) p120 and neurofibromin. Ras activation is buttressed by signals from integrin-cell-adhesion receptors, which activate Shc directly or by means of Src-family tyrosine kinases²⁹. In the GTP-bound state, Ras engages a complex network of effectors^{7,27,28}. Among these, the Ras–MEK–ERK kinase cascade has a key role in promoting CDK activation. ERK phosphorylates and stabilizes c-Myc, a transcription factor that induces the expression of cyclin D1 and suppresses that of CDK inhibitors⁷ (Fig. 2b).

The lipid kinase PI(3)K binds to receptor or IRS phosphotyrosine sites by means of SH2 domains in its non-catalytic α -subunit³⁰ (Fig. 3a). PI(3)K can also be recruited to the cell membrane by means of Ras. At the membrane, PI(3)K generates phosphatidylinositol-3, 4, 5-trisphosphate (PIP3) in a process that is undone by the lipid phosphatases PTEN and SHIP (Fig. 3a). PIP3-rich membrane locations attract the protein kinase Akt, also known as PKB, which then becomes activated by PDK kinases. Akt aids CDK activation by relieving two constraints (Fig. 3b): (1) Akt inhibits glycogen synthase kinase 3- β (GSK3- β), preventing this kinase from phosphorylating and destabilizing cyclin D (ref. 7); (2) Akt also inhibits FOXO transcription factors, barring them from the nucleus and thus from target genes that include p27Kip1 and p21Cip1/WAF1 (refs 30, 31).

Overexpression of mitogenic cytokines or constitutive activation of their receptors are common culprits in cancer. Examples include EGFR mutations in carcinomas of the lung, head and neck, colon, and in glioblastoma, HER2 (human EGFR-2) overexpression in breast cancer, PDGF (platelet-derived growth factor) overexpression in glioblastoma, and mutations in the PDGF and Kit receptors in gastrointestinal sarcomas^{32,33}. Ras-activating mutations are found in various types of carcinoma^{27,28}; activating B-Raf mutations are present in most melanomas³⁴, and activating mutations or amplification of PI(3)K and Akt occur with moderate frequency in a wide range of carcinomas³⁰. The frequent somatic loss of PTEN in glioblastoma, melanoma, lymphoma, myeloma and many carcinomas, adds to our growing view of the Ras and PI(3)K pathways as major targets of disruption in cancer^{27,33}.

Accordingly, a number of antibodies and small-molecule compounds targeting receptor tyrosine kinases^{32,33,35,36}, their ligands³⁷ or their degradative pathways³⁸ are entering clinical practice as new anti-cancer drugs. The humanized monoclonal antibody trastuzumab (Herceptin) against HER2 is effective in HER2-overexpressing breast cancer and is currently being used in trials for other cancers. Cetuximab (Erbix) raised against EGFR and bevacizumab (Avastin) against the angiogenic mitogen VEGF (vascular endothelial growth factor) are approved for the treatment of advanced and metastatic colon cancer. The small-molecule kinase inhibitor imatinib (Gleevec) is proving to be very effective against the cancer-causing fusion protein Bcr–Abl in chronic myeloid leukaemia and against mutant Kit in gastrointestinal sarcomas. Bcr–Abl is the product of the 'Philadelphia chromosome', a 9;22 chromosomal translocation that fuses the kinase domain of the non-receptor tyrosine kinase c-Abl to a portion of the *Bcr* gene product. Prompted by this success, other therapeutic tyrosine kinase inhibitors are emerging. Two that stand out for their effect in certain cases of lung cancer are gefitinib (Iressa) and erlotinib (Tarceva)^{39–41}.

Growth and survival functions

In addition to CDK activation, G1 progression requires a favourable metabolic state to accumulate enough cell mass and organelles to establish two daughter cells. G1 progression also depends on the cell's ability to avert developmental programmes or accidents that call for cell death. In the adult, a balance of cell death and compensatory proliferation maintains epithelia and the bone marrow in a state of constant renewal. Mature cells of the skin, intestinal lining or blood live only a few days before being eliminated by programmed death.

Programmed cell death is often triggered by a process of apoptosis, in which proteases known as caspases execute the orderly demise of

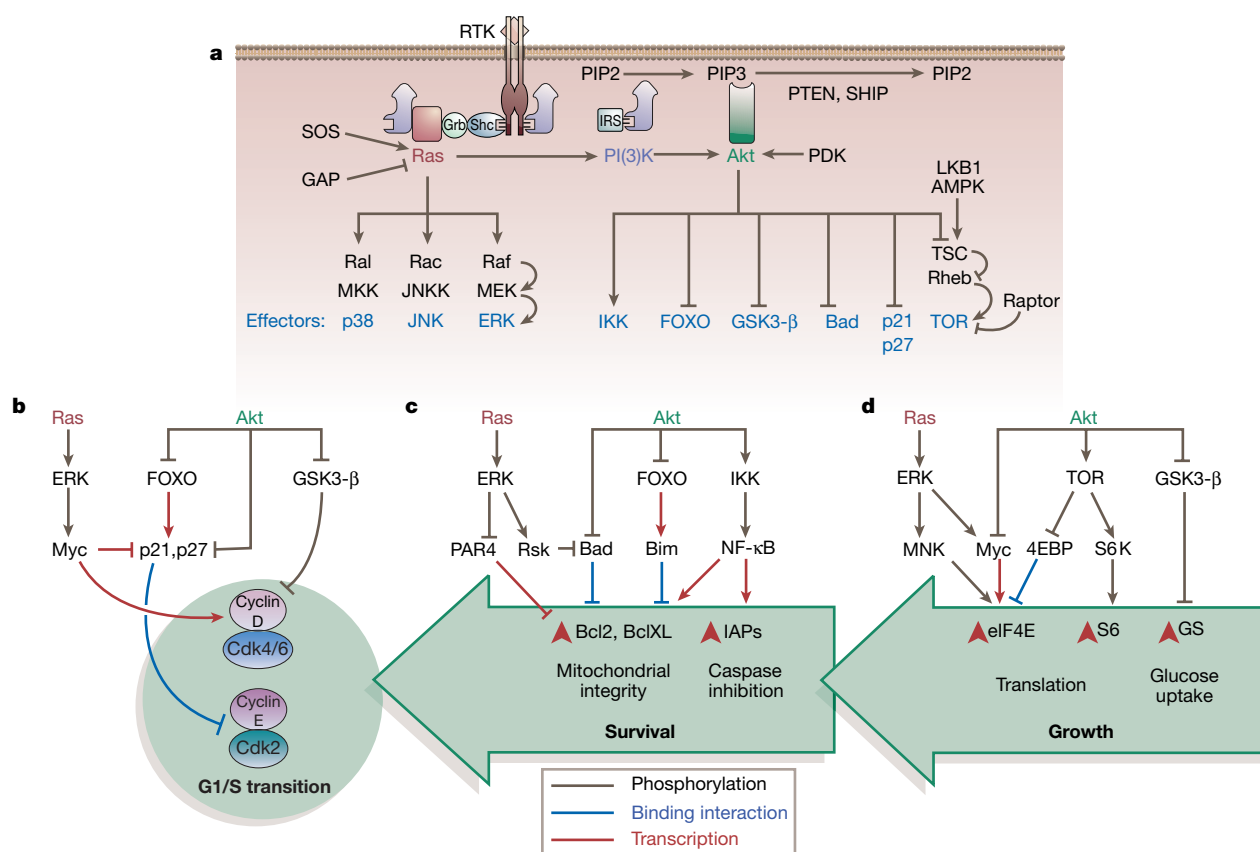


Figure 3 Networks integrating growth, survival and proliferation signals. **a**, Ras and PI(3)K pathways are activated by the recruitment of Ras and PI(3)K respectively to membrane assembly points around receptor tyrosine kinases (RTKs) and their phosphorylated substrates. RTKs generate phosphotyrosine sites (=) for the recruitment of Ras and PI(3)K. PI(3)K in turn generates PIP3 for the recruitment of Akt. Through several effectors, Ras and Akt coordinately control cell growth, survival

and proliferation functions. One of these effectors, TOR, also receives metabolic and nutrient inputs by means of AMPK and raptor. **b–d**, Ras and Akt effectors control cell growth, survival and proliferation through protein phosphorylation, binding interactions or transcriptional regulation. Among these effectors, the protein kinases ERK, TOR and GSK3-β, the transcription factors Myc and FOXO, and the CDK inhibitors p21 and p27, function as key nodes for signal integration.

the cell⁴² (see review in this issue by Lowe *et al.*, page 307). Caspases may be activated by extracellular factors, such as FasL (Fas ligand), which act by means of membrane death receptors^{42,43}. Caspases may also be activated by cell-intrinsic signals that release cytochrome *c* and the protein Smac/DIABLO from the mitochondria⁴⁴. In mammalian cells, the release of these factors is achieved by pro-apoptotic members of the Bcl2 gene family, including Bad, Bax, Bim, Noxa, Puma and others. These act in opposition to the anti-apoptotic members, BclXL, Mcl-1 and Bcl2 itself⁴². Cytochrome *c* drives the formation of the 'apoptosome', a complex with the signalling proteins caspase 9 and Apaf1, whereas Smac/DIABLO blocks inhibitors of apoptosis proteins (IAPs) — a family of caspase inhibitors. These effects achieve robust caspase activation. Other factors involved in this process and alternative mechanisms of cell death are also under investigation, which may be relevant to cancer^{42,43,45,46}. Cancer cells must avert death mechanisms to form a tumour. As first observed with the effect of Bcl2 on Myc-driven lymphomas in mice⁴⁷, a combination of mitogenic signals and excessive survival promotes the relentless expansion of tumour cell populations.

TOR, Myc and GSK3-β integrate growth and survival

Cell growth, survival and proliferation are distinct but intertwined functions whose integration is partly achieved by the versatile Ras and PI(3)K pathways. Along with CDK activation, these two pathways deliver potent growth and survival signals. PI(3)K supports cell survival by means of Akt-mediated post-translational and transcrip-

tional events³⁰. Akt phosphorylates and inhibits Bad (Fig. 3c) and prevents FOXO-dependent expression of Bim. Additional effects of Akt are mediated by the kinase IKK. IKK activates the transcription factor NF-κB to induce the expression of BclXL for the preservation of mitochondrial integrity, and IAPs for the inhibition of caspases. Moreover, IKK also phosphorylates FOXO, marking it for nuclear exit and destruction⁴⁸. For its part, Ras contributes to cell survival by means of ERK. In neurons, the ERK-activated kinase p90rsk phosphorylates and inhibits Bad⁴⁹. In fibroblasts, ERK inhibits PAR-4, a transcriptional repressor of *Bcl2* (ref. 50).

The serine protein kinase TOR (target of rapamycin) funnels diverse metabolic inputs and growth signals to the translational machinery⁵¹. On growth factor stimulation, Akt phosphorylates the tumour suppressor protein TSC2. It thereby turns off the GAP activity of a TSC1/2 complex, which is directed towards the G-protein Rheb. Rheb then activates TOR (Fig. 3a). Under conditions of ATP depletion, the protein kinase LKB1 activates the kinase AMPK, which then activates TSC2, leading to TOR inhibition^{52,53}. Furthermore, under amino-acid restriction, the protein raptor binds and inhibits TOR.

TOR has two effector arms for the activation of translation (Fig. 3d): (1) it phosphorylates S6K (ribosomal protein S6 kinase, also known as p70rsk) which stimulates translation elongation as well as the translation of ribosomal proteins⁵¹; (2) TOR also phosphorylates the eukaryotic initiation factor-4E (eIF-4E)-binding protein (4E-BP), suppressing its inhibitory effect on the translation-initiation

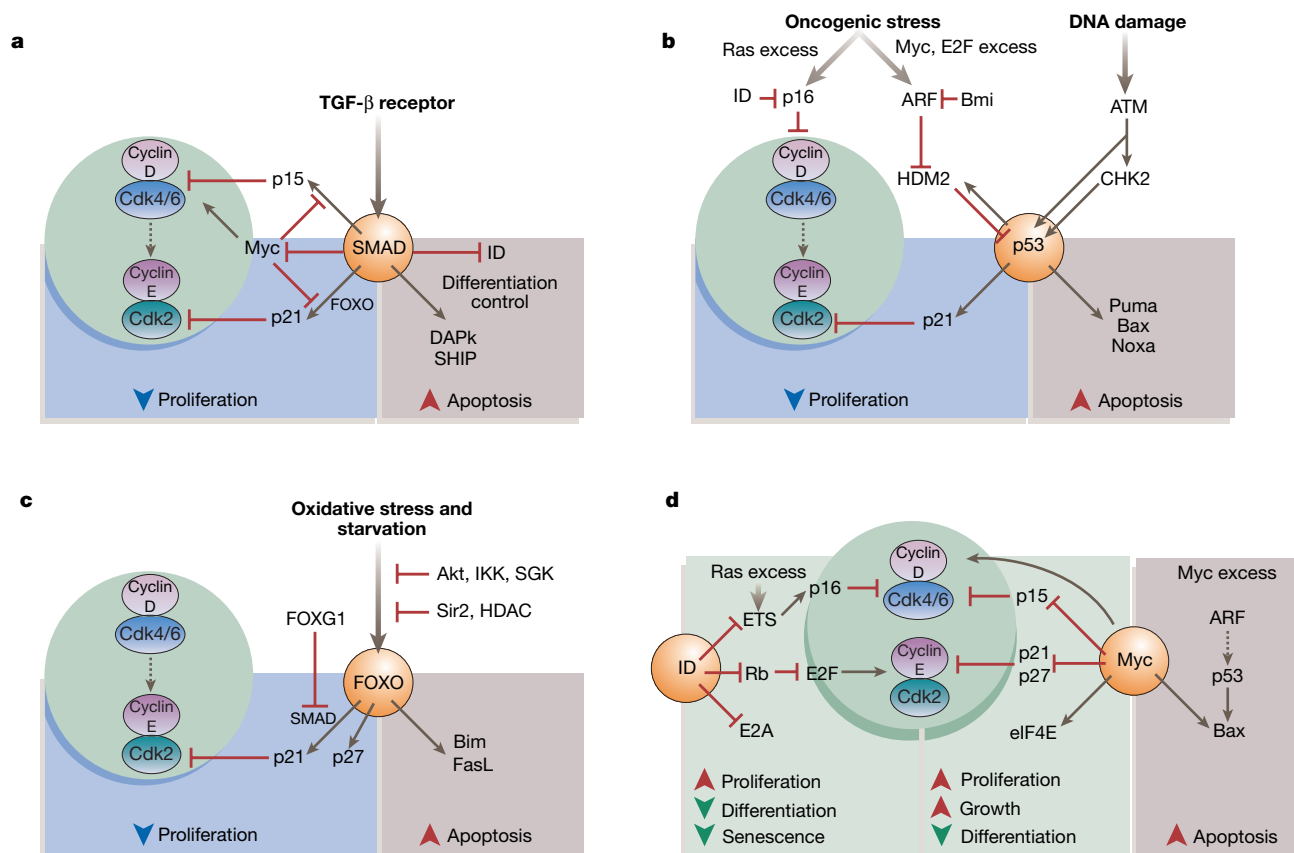


Figure 4 Signal integration by means of five transcriptional nodes. **a**, The SMAD node mediates cytositis (arrest of cell growth and multiplication) and apoptosis in response to TGF- β . SMADs and FOXO collaboratively activate the expression of cytositis genes. In contrast, SMAD and Myc work in opposition: SMADs inhibit Myc expression whereas Myc inhibits SMAD activation of cytositis genes. **b**, The p53 node mediates expression of cytositis and apoptotic factors in response to DNA damage or oncogenic levels of mitogenic signals. Intense Ras signalling can trigger either cytositis or cell senescence. In stem cells, ID and Bmi-1 prevent activation of this checkpoint to avert

premature exit from the proliferative state. **c**, The FOXO node mediates cytositis and apoptosis in response to oxidative stress (which causes FOXO acetylation) and starvation (which causes FOXO dephosphorylation). Several deacetylases (Sir2, HDAC) and growth-factor-activated kinases inhibit FOXO. **d**, ID and Myc suppress CDK inhibitors to favour cell proliferation. ID also sequesters E2A to prevent terminal differentiation. Myc induces eIF-4E to promote protein translation and cell growth. However, too much Myc activity triggers p53-dependent and p53-independent apoptosis, unless the p53 pathway is disabled, as occurs in most cancers.

factor eIF-4E (ref. 54). Within translation-initiation complexes, eIF-4E can be further activated by the ERK- or p38MAPK-dependent kinase MNK. The Ras-ERK and Akt-TOR inputs enhance ribosomal recruitment of different sets of messenger RNAs encoding cell-cycle components (for example, cyclin D), transcription factors (for example, c-Myc, ID), and cytokines (for example, VEGF; FGF, for fibroblast growth factor)⁵⁵. eIF-4E also receives inputs by means of Myc. These inputs activate the expression of eIF-4E as well as other translation initiators. eIF-4E cooperates with c-Myc in B-cell lymphomagenesis in the mouse^{56,57}, supporting the idea that eIF-4E may be involved in cancer⁵⁴. Mutations that inactivate *TSC2* or *LKB1* also predispose organisms to cancer. Therefore, the TOR-eIF-4B pathway is a major conduit for external and internal signals controlling translation. It is also the object of increasing attention as a possible therapeutic target. An interesting drug in this respect is the TOR blocker rapamycin, which is already in clinical use as an immunosuppressant⁵¹.

Completing the Akt network is Akt's inhibitory effect on yet another multifaceted kinase, GSK3- β , whose substrates include various regulators of cell proliferation and metabolism^{58,59} (Fig. 3b, d). GSK3- β phosphorylates and inactivates cyclin D, glycogen synthase, several transcription factors (for example, c-Myc) and mitogenic signal transducers (for example, β -catenin and GLi; see section 'G1 inputs in stem cells' below). When phosphorylated by Akt, GSK3- β

cannot recognize c-Myc or glycogen synthase, which are therefore spared from inactivation.

The transcription factors c-Myc, L-Myc and N-Myc all activate transcription by dimerizing with the related protein Max to induce the expression of D-type cyclins⁶⁰. In association with Miz1 (Myc interacting zinc-finger protein), c-Myc inhibits the expression of CDK inhibitors p21Cip1 and p15Ink4b (refs 61, 62). In some settings, the effects of Myc on cell growth are more pronounced than its proliferative action⁶⁰. Furthermore, Myc affects terminal differentiation and — when its high activity becomes an oncogenic risk — it triggers apoptosis⁶³ (see section 'DNA damage and stress checkpoints' below; Fig. 4d). When apoptosis is averted, Myc becomes oncogenic^{47,63}. The overexpression of c-Myc is observed in many tumours and is a causative event in Burkitt's lymphoma. Thus, Myc constitutes a key node for the management of growth, proliferation and differentiation signals as well as oncogenic stress.

Cytostatic signalling by means of the SMAD node

G1 progression and cell proliferation are limited by extracellular signals that maintain tissue homeostasis. The cytokine TGF- β (transforming growth factor- β) and its family members activin and myostatin are potent suppliers of such signals⁶⁴. TGF- β signalling promotes growth and development during early embryogenesis and in some adult mesenchymal cells. In mature tissues, however, many

cells respond to TGF- β with cytotaxis and, in some cases, with apoptosis. Epithelial, endothelial, haematopoietic and neural progenitor cells, as well as certain mesenchymal cell types are programmed for these types of TGF- β responses.

TGF- β activates a membrane-receptor serine/threonine kinase complex that phosphorylates the tumour suppressor proteins Smad2 and Smad3 (ref. 65). Activated SMAD proteins accumulate in the nucleus and form transcriptional complexes with the tumour suppressor protein Smad4 and different DNA-binding partners, co-activators and co-repressors. These complexes target different genes depending on their composition, and collectively they generate hundreds of cell-specific gene responses, some of which are devoted to arresting G1. The TGF- β cytotaxis programme in epithelial cells involves the induction of p21Cip1 and p15Ink4b expression and the repression of Myc and ID (ref. 64; Fig. 4a). Like Myc, ID factors support cell proliferation (see section 'Averting differentiation with ID and Bmi-1' below). Once p21Cip1 is relieved from Myc-mediated repression, it is transactivated by a SMAD-FOXO complex⁶⁶. FOXO transcription factors (FOXO-1, -3, -4 and -6) are members of the Forkhead family that constitute a major regulatory node in the control of cell and organismal growth, development, metabolism and longevity under caloric restriction^{31,67}. The activity of FOXO factors as SMAD partners is inhibited by Akt and by binding of the telencephalic development factor FOXG1 (ref. 66). The induction of the death-associated protein kinase DAPk and the inositol-5-phosphatase SHIP, which limits Akt survival signalling, have been implicated in TGF- β -mediated apoptosis⁵⁸.

Cancer cells may exploit any opportunity to lose cytotaxis responsiveness to TGF- β (refs 64, 68, 69). Indeed, inactivating mutations in the TGF- β type-II receptor are very common in colon cancers with defects in replication-error repair. Inactivation of *Smad4* (also known as *DPC4*, deleted in pancreatic carcinoma locus 4) occurs in one-half of pancreatic carcinomas. But more often, tumour cells lose cytotaxis responsiveness because of defects downstream of SMAD. In such cases, tumour cells may use other aspects of TGF- β signalling, thereby exacerbating their own proliferative, invasive and metastatic behaviour. TGF- β can thus switch from a tumour suppressor into a tumour instigator^{64,68}.

DNA damage and stress checkpoints

Replication errors and DNA damage by radiation or chemical agents constantly challenge the genetic integrity of a cell. In defence, multi-protein systems detect such mishaps and stall G1 or G2 until the damage is repaired, or else trigger the elimination of the cell by apoptosis. When such quality control 'checkpoints' fail, cells accumulate genomic defects that are conducive to cancer (see review in this issue by Kastan and Bartek, page 316). Indeed, one of the key players in these pathways, the transcription factor p53, is the most frequently mutated tumour suppressor gene in human cancer^{11,70}. Cells lacking p53 can survive levels of DNA damage and mitogenic stimulation that would otherwise kill the cell.

DNA damage is sensed by protein assemblies whose effector components are the protein kinases ATM (ataxia telangiectasia mutated) and ATR (ATM-related)⁷¹. ATM recognizes double-strand breaks caused by ionizing radiation or other factors (Fig. 2). ATR is activated by ultraviolet irradiation (possibly expressed by DNA-base damage) and by unreplicated DNA. ATM and ATR activate the transducer checkpoint kinases CHK2 and CHK1, respectively. CHK1 inhibits CDC25 phosphatases by causing their nuclear exclusion or degradation, which in turn prompts the accumulation of inhibitory phosphorylation of CDKs. CHK2 and ATM signal through p53, which induces the expression of p21Cip1/WAF1, various pro-apoptotic factors (Puma, Bax, Noxa), DNA-repair and oxidative-stress-response genes, and the feedback regulator HDM2 (the human orthologue of the mouse double minute 2, MDM2)(ref. 72; Fig. 4b). For activation of pro-apoptotic genes, p53 requires the collaboration of its family members, p63 and p73 (ref. 73), although this may not

apply to all cell types⁷⁴. Apoptosis is favoured over cytotaxis by the protein ASPP, which facilitates the activation of pro-apoptotic p53 target genes⁷², and by Myc, which suppresses p21Cip1 expression and cytotaxis^{60,62}.

p53 is a short-lived protein that is stabilized and transcriptionally activated by ATM-mediated phosphorylation^{71,72}. Acetylation of p53 further enhances its transcriptional activity, whereas the deacetylase Sir2 negatively regulates p53 and favours cell survival under stress⁷⁵. The ubiquitin ligase HDM2 binds p53 causing its ubiquitination and subsequently its nuclear exclusion or destruction through mechanisms that are still a matter of debate. Another G1 checkpoint function served by p53 and Rb is to supervise hyperactive Ras, Myc and E2F signalling⁷⁶. These inputs are channelled by means of the *Ink4* locus, which encodes two distinct tumour suppressor products p16Ink4a and p14ARF (p19ARF in the mouse; Fig. 4b). p16Ink4a expression is induced by ETS transcription factors in response to excess Ras-ERK activity. p16Ink4a protects the cell against hyperactive Ras-ERK by blocking Cdk4/6 and thereby inducing cell quiescence or senescence. ARF expression is somehow induced by many oncogenic signals, such as those conveyed by excessive Ras, Myc or E2F activity. ARF inhibits HDM2, allowing p53 activation. Like p53, the *Ink4* locus is also frequently altered in cancer by mutations that inactivate one or both of its products⁷⁶.

The ability of cells to undertake the G1/S transition is also restricted by starvation and oxidative stress. These inputs are mediated in part through FOXO factors^{31,67} (Fig. 4c). Under starvation, FOXO factors accumulate in the nucleus to induce the expression of cytotaxis (p27Kip1) and pro-apoptotic factors (Bim, FasL). Under conditions of oxidative stress, FOXO becomes acetylated and maximally active^{77,78}. The deacetylases Sir2 and HDAC counter this effect and suppress the ability of FOXO to activate pro-apoptotic genes. On activation by growth and survival signals, Akt directly or by means of IKK, as well as SGK (serum and glucocorticoid-activated kinase), phosphorylate FOXO factors at multiple sites^{31,48}. This forces FOXO out of the nucleus and, at least in the case of IKK, marks FOXO for destruction. Thus, FOXO proteins are maximally active under oxidative stress and starvation, and may be switched by deacetylation into promoting endurance (that is, cell quiescence without death). By acting as partners of SMAD transcription factors in the induction of p21Cip1 expression⁶⁶, FOXO proteins additionally provide a link to TGF- β cytotaxis signals.

There is increasing evidence for the involvement of FOXO factors in cancer⁷⁹. Decreased accumulation of FOXO proteins in the nucleus owing to the presence of hyperactive Akt or IKK correlates with poor survival in breast cancer⁴⁸. Moreover, FOXO2 and FOXO4 are disrupted by chromosomal translocations in a subset of acute leukaemias⁸⁰, and FOXO1 in a subtype of rhabdomyosarcoma⁷⁹.

Averting differentiation with ID and Bmi-1

After leaving the stem-cell stage, cells inexorably differentiate as they divide, and eventually undergo terminal differentiation or death. Averting premature differentiation is crucial during embryogenesis and for tissue homeostasis in the adult. This role partly falls on members of the ID family (ID1 to 4)⁸¹. ID proteins sequester the transcription factor E2A away from transcriptional inducers of differentiation such as MyoD and myogenin, which must dimerize with E2A for transactivation (Fig. 4d). ID proteins inhibit the terminal differentiation of lineages as diverse as myoblasts, angioblasts, oligodendrocytes, lymphocytes and dendritic cells. ID factors also inhibit earlier differentiation steps: induction of ID1 and ID3 by BMP4 (bone morphogenetic protein 4) supports self-renewal of embryonic stem cells⁸². Moreover, ID factors establish inhibitory interactions with Rb (ref. 81). The differentiation defects of mice lacking Rb largely disappear in an ID2-defective background. Another role of ID proteins is to protect the cell against Ras-induced senescence: ID prevents ETS factors from inducing p16Ink4a expression⁸¹ (Fig. 4b,d).

Transcription from the *Ink4* locus can also be inhibited by Bmi-1, a component of Polycomb transcriptional-silencing complexes⁸³. Bmi-1 is expressed in haematopoietic stem cells (HSCs) and lymphoid progenitor cells. By preventing ARF-mediated apoptosis under mitogenic stimulation, Bmi-1 helps to maintain HSC populations and allows the expansion of leukaemias and Myc-dependent lymphomas in the mouse^{84,85}. Differentiation of HSCs is also inhibited by Myc. In the skin, however, c-Myc seems to have the opposite effect, as it depletes the stem cell pool by driving proliferation and differentiation along the keratinocyte lineage⁸⁶.

G1 inputs in stem cells

Cancer most often emerges in those tissues that are undergoing constant renewal by the proliferation of stem cells. Mutations that expand the numbers and misguide the fate of these cells are perpetuated to their entire progeny, increasing the chances of additional mutations and tumour progression. By retaining or re-acquiring stem-cell behaviour, malignant cells may be able to recreate a tumour during metastasis or after therapeutic elimination of the bulk of the tumour population⁸⁷. This provides a rationale for targeting tumour stem cells being the most effective way to treat cancer. Although it may not apply to all cases, this hypothesis is supported by persuasive evidence in leukaemia⁸⁸ and breast cancer⁸⁹ (see also review in this issue by Beachy *et al.*, page 324).

These ideas have drawn attention to the pathways that control stem-cell proliferation. Among these, the Wnt and Hedgehog pathways are of particular relevance to cancer. Both pathways feature prominently in cell-fate specification and pattern formation during embryogenesis and adult tissue renewal, and they show striking parallels in their overall traits. Both pathways are set off by short-range acting ligands that trigger dual-membrane protein interactions. In both cases, these interactions activate latent transcription factors by suppressing proteolytic events. The proliferative effect of both pathways involves the transcriptional induction of Myc.

Wnt signalling inhibits GSK3- β preventing the complex GSK3- β -Axin-APC (adenomatous polyposis coli) from phosphorylating β -catenin and targeting it for destruction in the cytoplasm. As a result, β -catenin moves into the nucleus and binds TCF/LEF (T-cell factor/lymphoid enhancer factor), turning these factors from transcriptional repressors into activators. The role of Wnt in promoting stem cell proliferation is particularly evident in the intestinal epithelium. Located in the intestinal crypts, stem cells constantly generate progeny that differentiate as they flow upward to the tip of the villi, where they die within days. TCF-mediated induction of c-Myc, with secondary induction of cyclin D1 and downregulation of p21Cip1, is thought to drive proliferation in these cells and their malignant derivatives⁹⁰.

Sonic hedgehog (Shh) and its family members Indian and Desert hedgehog bind to one multi-pass membrane protein, Patched (Ptch), to lift its inhibitory effect on another, Smoothened (Smo)⁹¹. This leads to the activation of Gli zinc-finger transcription factors. The Shh transcription programme for G1 progression has been studied in granule-cell precursors (GCPs). These cells give rise to a major class of neurons, and to medulloblastoma in children. The most robust of these gene responses is *N-Myc* induction, which in turn promotes the expression of cyclin D1 and *E2F1*, thereby stimulating cell proliferation^{92,93}.

Mutations that constitutively activate the Wnt pathway are essentially universal in colon cancer⁹⁴. Inherited mutations that inactivate APC are responsible for familial adenomatous polyposis, a highly penetrant cancer predisposition syndrome. Inactivation of both APC alleles is observed in most colon cancers. Tumours with wild-type APC are often defective in Axin or contain degradation-resistant β -catenin mutants. A constitutively active Wnt/ β -catenin pathway increases the pool of epithelial progenitors in the proliferative state. This gives rise to overgrowths (polyps) that are primed for additional mutations and progression to carcinoma. A role for Wnt signalling in human breast cancer is possible. Chronic myeloid leukaemia (CML)

progression in humans occurs with β -catenin activation⁸⁸. Thus, Wnt signalling preferentially targets stem/progenitor cells or forces differentiated cells to re-acquire this phenotype. As for Hedgehog, hereditary or somatic mutations that inactivate Ptch or activate Smo cause basal-cell skin carcinoma, medulloblastoma and rhabdomyosarcoma⁹¹. Activation of this pathway by overexpressed ligands occurs in carcinomas of the stomach, pancreas, bile duct or oesophagus^{95–97}, drawing further attention to its relevance in cancer.

G1 signalling specificity

By focusing on the principal networks that control G1 progression, the preceding sections have overlooked a multitude of variant pathways that provide crucial regulatory flexibility in a cell-type-specific manner. Most of the signalling components discussed above exist in several versions that are encoded by paralogous genes or that are the products of alternative splicing of the same transcript. This variegation provides cell-type specificity as well as redundancy and robustness to G1 pathways. With this also comes tissue-restricted fragilities. Components that operate in many cell types are found to contribute to oncogenesis in only a few. For example, Ras signalling operates in virtually all cell types, but Ras is mutated to an oncogenic form in only about 15% of human cancers^{27,34}. *K-Ras* mutations are frequent and *H-Ras* mutations rare in tumours of the lung or pancreas, whereas the opposite seems to be true in bladder cancer. *N-Ras* mutations are present in melanoma and hepatocellular carcinoma—tumours that never seem to have *K-Ras* or *H-Ras* mutations. Ras mutations of any kind are rare in primary breast tumours, even though HER2 amplification that activates Ras signalling is found in these tumours. The case of Ras typifies that shown by many G1 regulators involved in cancer.

Several factors may account for the tissue-specific potential of each cancer gene, and identifying such factors remains an important challenge. Any given cancer gene may be more susceptible than others to the mutagens that a particular tissue is exposed to. Also, redundant gene-family members may compensate for the inactivating of a tumour suppressor. For example, most tissues may be less susceptible to *Rb* mutations than the developing retina because *p107* and *p130* may compensate for *Rb* loss in those tissues⁹. In the case of dominant oncogenes, an activating mutation will be tumorigenic to the extent that it can feed into an effector pathway and provide a survival or proliferative advantage. And, once the first tumorigenic event is in place, the additional mutations that must occur for tumour progression may depend on the nature of that initiating event as well as the developmental stage of the cell that sustained it. For example, colon cancers initiated by a germline mutation in APC progress differently from colon cancers initiated by germline mutation in genes that repair replication errors⁹⁴.

Outlook

Ascertaining what dictates the oncogenic potential of G1 regulatory components is important for improving cancer therapy. By evading proper behaviour, malignant cells may evolve into a stage where they are critically dependent on certain survival mechanisms that are dispensable in normal tissue. This would open an interesting window for therapeutic intervention. CML, which is driven by Bcr-Abl, provides a case in point. By acting as a mislocalized kinase, Bcr-Abl may phosphorylate substrates that c-Abl does not normally touch. Thus, myeloid progenitor cells survive in a proliferative state and behave as leukaemia stem cells because of an idiosyncratic signal provided by Bcr-Abl (ref. 88). Blocking that signal would be catastrophic for the survival of CML cells, which is precisely what the acclaimed Abl kinase blocker imatinib (Gleevec) accomplishes.

Normally, however, things are more complicated^{33,36}. The fact that cancer results from the combined action of multiple oncogenic alterations argues that single-agent therapies will remain the exception. Most drugs targeting G1 components may have to be used in tumour-specific combinations. Recent experience supports this.

The EGFR kinase inhibitors gefitinib and erlotinib are effective in non-small-cell lung carcinomas (NSCLCs) that contain mutant EGFR^{39–41} but less so in glioblastomas that suffer from EGFR activation as a late event in their progression. Glioblastomas frequently harbour an activated PI(3)K pathway⁴⁰, whereas oncogenic activation of this pathway is infrequent in NSCLC. Perhaps combinations of EGFR blockers and PI(3)K blockers would be effective in glioblastoma. Even in the case of imatinib, combination therapy may be in order. The effectiveness of this drug in CML drops when the disease accumulates additional mutations and enters an acute phase. Moreover, imatinib therapy is not free from the emergence of resistance, and may have to be combined with other Bcr–Abl inhibitors in long-term treatments⁹⁸.

These challenges notwithstanding, the growing ability to develop new therapies and explain why they may work or fail indicates the progress achieved. The core machinery of the cell cycle has been delineated, at least in broad strokes. Indeed, the results emerging from our plunge into the bewildering complexity of cell-cycle signalling networks suggest that we may grasp their logic soon. In particular, the rapid acquisition of data from new high-throughput technologies promises a better understanding of normal and altered pathways in both healthy and cancerous cells. Many questions remain and many principles need to be established, but a problem that has long fascinated biologists is beginning to be unravelled. □

doi:10.1038/nature03094

1. Morgan, D. O. Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell Dev. Biol.* **13**, 261–291 (1997).
2. Murray, A. W. Recycling the cell cycle: cyclins revisited. *Cell* **116**, 221–234 (2004).
3. Kelly, T. J. & Brown, G. W. Regulation of chromosome replication. *Annu. Rev. Biochem.* **69**, 829–880 (2000).
4. Prasanth, S. G., Mendez, J., Prasanth, K. V. & Stillman, B. Dynamics of pre-replication complex proteins during the cell division cycle. *Phil. Trans. R. Soc. Lond. B* **359**, 7–16 (2004).
5. Geng, Y. *et al.* Cyclin E ablation in the mouse. *Cell* **114**, 431–443 (2003).
6. Ortega, S. *et al.* Cyclin-dependent kinase 2 is essential for meiosis but not for mitotic cell division in mice. *Nature Genet.* **35**, 25–31 (2003).
7. Sears, R. C. & Nevins, J. R. Signaling networks that link cell proliferation and cell fate. *J. Biol. Chem.* **277**, 11617–11620 (2002).
8. Stevaux, O. & Dyson, N. J. A revised picture of the E2F transcriptional network and RB function. *Curr. Opin. Cell Biol.* **14**, 684–691 (2002).
9. Lipinski, M. M. & Jacks, T. The retinoblastoma gene family in differentiation and development. *Oncogene* **18**, 7873–7882 (1999).
10. Sherr, C. J. & Roberts, J. M. CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev.* **13**, 1501–1512 (1999).
11. Sherr, C. J. Principles of tumor suppression. *Cell* **116**, 235–246 (2004).
12. Geng, Y. *et al.* Rescue of cyclin D1 deficiency by knockin cyclin E. *Cell* **97**, 767–777 (1999).
13. Kozar, K. *et al.* Mouse development and cell proliferation in the absence of d-cyclins. *Cell* **118**, 477–491, (2004).
14. Malumbres M. *et al.* Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell* **118**, 493–504, (2004).
15. Pavletich, N. P. Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and Ink4 inhibitors. *J. Mol. Biol.* **287**, 821–828 (1999).
16. Blain, S. W., Scher, H. I., Cordon-Cardo, C. & Koff, A. p27 as a target for cancer therapeutics. *Cancer Cell* **3**, 111–115 (2003).
17. Montagnoli, A. *et al.* Ubiquitination of p27 is regulated by Cdk-dependent phosphorylation and trimeric complex formation. *Genes Dev.* **13**, 1181–1189 (1999).
18. Reed, S. I. Ratchets and clocks: the cell cycle, ubiquitylation and protein turnover. *Nature Rev. Mol. Cell Biol.* **4**, 855–864 (2003).
19. Bashir, T., Dorrello, N. V., Amador, V., Guardavaccaro, D. & Pagano, M. Control of the SCF (Skp2-Cks1) ubiquitin ligase by the APC/C(Cdh1) ubiquitin ligase. *Nature* **428**, 190–193 (2004).
20. Wei, W. *et al.* Degradation of the SCF component Skp2 in cell-cycle phase G1 by the anaphase-promoting complex. *Nature* **428**, 194–198 (2004).
21. Petronczki, M., Siomos, M. F. & Nasmyth, K. Un menage a quatre: the molecular biology of chromosome segregation in meiosis. *Cell* **112**, 423–440 (2003).
22. Rape, M. & Kirschner, M. W. Autonomous regulation of the anaphase-promoting complex couples mitosis to S-phase entry. *Nature* (in the press).
23. Hsu, J. Y. *et al.* E2F-dependent accumulation of hEmi1 regulates S phase entry by inhibiting APC (Cdh1). *Nature Cell Biol.* **4**, 358–366 (2002).
24. Yu, Q., Geng, Y. & Sicinski, P. Specific protection against breast cancers by cyclin D1 ablation. *Nature* **411**, 1017–1021 (2001).
25. Pawson, T. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116**, 191–203 (2004).
26. Schlessinger, J. & Lemmon, M. A. SH2 and PTB domains in tyrosine kinase signaling. *Sci. STKE* **191**, RE12 (2003).
27. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature Rev. Cancer* **3**, 11–22 (2003).
28. Coleman, M. L., Marshall, C. J. & Olson, M. F. RAS and RHO GTPases in G1-phase cell-cycle regulation. *Nature Rev. Mol. Cell Biol.* **5**, 355–366 (2004).
29. Giancotti, F. G. & Tarone, G. Positional control of cell fate through joint integrin/receptor protein kinase signaling. *Annu. Rev. Cell Dev. Biol.* **19**, 173–206 (2003).
30. Vivanco, I. & Sawyers, C. L. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nature Rev. Cancer* **2**, 489–501 (2002).
31. Tran, H., Brunet, A., Griffith, E. C. & Greenberg, M. E. The many forks in FOXO's road. *Sci. STKE* **172**, RE5 (2003).
32. Gschwind, A., Fischer, O. M. & Ullrich, A. The discovery of receptor tyrosine kinases: targets for cancer therapy. *Nature Rev. Cancer* **4**, 361–370 (2004).
33. Sawyers, C. L. Opportunities and challenges in the development of kinase inhibitor therapy for cancer. *Genes Dev.* **17**, 2998–3010 (2003).
34. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
35. Noble, M. E., Endicott, J. A. & Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **303**, 1800–1805 (2004).
36. Arteaga, C. L. & Baselga, J. Tyrosine kinase inhibitors: why does the current process of clinical development not apply to them? *Cancer Cell* **5**, 525–531 (2004).
37. Ferrara, N., Hillan, K. J., Gerber, H. P. & Novotny, W. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. *Nature Rev. Drug Discov.* **3**, 391–400 (2004).
38. Chiosis, G. L. B., Huez, H., Solit, D., Basso, A. & Rosen, N. Development of purine-scaffold small molecule inhibitors of Hsp90. *Curr. Cancer Drug Targets* **3**, 371–376 (2003).
39. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
40. Maher, E. A. *et al.* Malignant glioma: genetics and biology of a grave matter. *Genes Dev.* **15**, 1311–1333 (2001).
41. Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib (Iressa®) and erlotinib (Tarceva TM). *Proc. Natl Acad. Sci. USA* **101**, 13306–13311 (2004).
42. Danial, N. N. & Korsmeyer, S. J. Cell death: critical control points. *Cell* **116**, 205–219 (2004).
43. Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* **281**, 1305–1308 (1998).
44. Jiang, X. & Wang, X. Cytochrome C-mediated apoptosis. *Ann. Rev. Biochem.* **73**, 87–106 (2004).
45. Davis, R. J. Signal transduction by the JNK group of MAP kinases. *Cell* **103**, 239–252 (2000).
46. Kolesnick, R. & Fuks, Z. Radiation and ceramide-induced apoptosis. *Oncogene* **22**, 5897–5906 (2003).
47. Cory, S. & Adams, J. M. The Bcl2 family: regulators of the cellular life-or-death switch. *Nature Rev. Cancer* **2**, 647–656 (2002).
48. Hu, M. C. *et al.* IkappaB kinase promotes tumorigenesis through inhibition of forkhead FOXO3a. *Cell* **117**, 225–237 (2004).
49. Bonni, A. *et al.* Cell survival promoted by the Ras-MAPK signaling pathway by transcription-dependent and -independent mechanisms. *Science* **286**, 1358–1362 (1999).
50. Barradas, M., Monjas, A., Diaz-Meco, M. T., Serrano, M. & Moscat, J. The downregulation of the pro-apoptotic protein Par-4 is critical for Ras-induced survival and tumor progression. *EMBO J.* **18**, 6362–6369 (1999).
51. Bjornsti, M. A. & Houghton, P. J. The TOR pathway: a target for cancer therapy. *Nature Rev. Cancer* **4**, 335–348 (2004).
52. Lizcano, J. M. *et al.* LKB1 is a master kinase that activates 13 kinases of the AMPK subfamily, including MARK/PAR-1. *EMBO J.* **23**, 833–843 (2004).
53. Shaw, R. J. *et al.* The LKB1 tumor suppressor negatively regulates mTOR signaling. *Cancer Cell* **6**, 91–99 (2004).
54. Holland, E. C., Sonenberg, N., Pandolfi, P. P. & Thomas, G. Signaling control of mRNA translation in cancer pathogenesis. *Oncogene* **23**, 3138–3144 (2004).
55. Rajasekhar, V. K. *et al.* Oncogenic Ras and Akt signaling contribute to glioblastoma formation by differential recruitment of existing mRNAs to polysomes. *Mol. Cell* **12**, 889–901 (2003).
56. Wendel, E. J. Survival signaling by Akt and eIF4E in oncogenes and cancer therapy. *Nature* **428**, 332–337 (2004).
57. Ruggero, D. *et al.* The translation factor eIF-4E promotes tumor formation and cooperates with c-Myc in lymphomagenesis. *Nature Med.* **10**, 484–486 (2004).
58. Cohen, P. & Frame, S. The renaissance of GSK3. *Nature Rev. Mol. Cell Biol.* **2**, 769–776 (2001).
59. Jope, R. S. & Johnson, G. V. The glamour and gloom of glycogen synthase kinase-3. *Trends Biochem. Sci.* **29**, 95–102 (2004).
60. Grandori, C., Cowley, S. M., James, L. P. & Eisenman, R. N. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16**, 653–699 (2000).
61. Wanzel, M., Herold, S. & Eilers, M. Transcriptional repression by Myc. *Trends Cell Biol.* **13**, 146–150 (2003).
62. Seoane, J., Le, H. V. & Massagué, J. Myc suppression of the p21 (Cip1) Cdk inhibitor influences the outcome of the p53 response to DNA damage. *Nature* **419**, 729–734 (2002).
63. Pelengaris, S., Khan, M. & Evan, G. c-MYC: more than just a matter of life and death. *Nature Rev. Cancer* **2**, 764–776 (2002).
64. Siegel, P. M. & Massagué, J. Cytostatic and apoptotic actions of TGF-beta in homeostasis and cancer. *Nature Rev. Cancer* **3**, 807–821 (2003).
65. Shi, Y. & Massagué, J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell* **113**, 685–700 (2003).
66. Seoane, J., Le, H. V., Shen, L., Anderson, S. A. & Massagué, J. Integration of Smad and forkhead pathways in the control of neuroepithelial and glioblastoma cell proliferation. *Cell* **117**, 211–223 (2004).
67. Accili, D. & Arden, K. C. FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell* **117**, 421–426 (2004).
68. Derynck, R., Akhurst, R. J. & Balmain, A. TGF-beta signaling in tumor suppression and cancer progression. *Nature Genet.* **29**, 117–129 (2001).
69. Roberts, A. B. & Wakefield, L. M. The two faces of transforming growth factor beta in carcinogenesis. *Proc. Natl Acad. Sci. USA* **100**, 8621–8623 (2003).
70. Michael, D. & Oren, M. The p53 and Mdm2 families in cancer. *Curr. Opin. Genet. Dev.* **12**, 53–59 (2002).
71. Sancar, A., Lindsey-Boltz, L. A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem.* **73**, 39–85 (2004).
72. Vousden, K. H. & Lu, X. Live or let die: the cell's response to p53. *Nature Rev. Cancer* **2**, 594–604 (2002).
73. Flores, E. *et al.* p63 and p73 are required for p53-dependent apoptosis in response to DNA damage.

- Nature* **416**, 560–564 (2002).
74. Senoo, M., Manis, J. P., Alt, F. W. & McKeon, F. p63 and p73 are not required for the development and p53-dependent apoptosis of T cells. *Cancer Cell* **6**, 85–89 (2004).
75. Brooks, C. L. & Gu, W. Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr. Opin. Cell Biol.* **15**, 164–171 (2003).
76. Lowe, S. W. & Sherr, C. J. Tumor suppression by Ink4a-Arf: progress and puzzles. *Curr. Opin. Genet. Dev.* **13**, 77–83 (2003).
77. Brunet, A. *et al.* Stress-dependent regulation of FOXO transcription factors by the SIRT1 deacetylase. *Science* **303**, 2011–2015 (2004).
78. Motta, M. C. *et al.* Mammalian SIRT1 represses forkhead transcription factors. *Cell* **116**, 551–563 (2004).
79. Arden, K. C. FoxO: linking new signaling pathways. *Mol. Cell* **14**, 416–418 (2004).
80. So, C. W. & Cleary, M. L. MLL-AFX requires the transcriptional effector domains of AFX to transform myeloid progenitors and transdominantly interfere with forkhead protein function. *Mol. Cell Biol.* **22**, 6542–6552 (2003).
81. Ruzinova, M. B. & Benezra, R. Id proteins in development, cell cycle and cancer. *Trends Cell Biol.* **13**, 410–418 (2003).
82. Ying, Q. L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**, 281–292 (2003).
83. Jacobs, J. J. *et al.* Bmi-1 collaborates with c-Myc in tumorigenesis by inhibiting c-Myc-induced apoptosis via INK4a/ARF. *Genes Dev.* **13**, 2678–2690 (1999).
84. Park, I. K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423**, 302–305 (2003).
85. Lessard, J. & Sauvageau, G. Bmi-1 determines the proliferative capacity of normal and leukaemic stem cells. *Nature* **423**, 255–260 (2003).
86. Owens, D. M. & Watt, F. M. Contribution of stem cells and differentiated cells to epidermal tumours. *Nature Rev. Cancer* **3**, 444–451 (2003).
87. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
88. Jamieson, C. H. *et al.* Granulocyte-macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.* **351**, 657–667 (2004).
89. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
90. van de Wetering, M. *et al.* The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111**, 241–250 (2002).
91. Lum, L. & Beachy, P. A. The Hedgehog response network: sensors, switches, and routers. *Science* **304**, 1755–1759 (2004).
92. Oliver, T. G. *et al.* Transcriptional profiling of the Sonic hedgehog response: a critical role for N-myc in proliferation of neuronal precursors. *Proc. Natl Acad. Sci. USA* **100**, 7331–7336 (2003).
93. Kerney, A. M., Cole, H. D. & Rowitch, D. H. Nmyc upregulation by sonic hedgehog signalling promotes proliferation in developing cerebellar granule neurone precursors. *Development* **130**, 15–28 (2003).
94. Grady, W. M. & Markowitz, S. D. Genetic and epigenetic alterations in colon cancer. *Annu. Rev. Genomics Hum. Genet.* **3**, 101–128 (2002).
95. Berman, D. M. *et al.* Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours. *Nature* **425**, 846–851 (2003).
96. Thayer, S. P. *et al.* Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. *Nature* **425**, 851–856 (2003).
97. Karhadkar, S. S. Hedgehog signalling in prostate regeneration, neoplasia and metastasis. *Nature* **431**, 707–712 (2004).
98. Shah, N. P. *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399–401 (2004).

Acknowledgements The author thanks E. Holland, N. Rosen and D. Solit for helpful discussions.

Competing interests statement The author declares that he has no competing financial interests.

Intrinsic tumour suppression

Scott W. Lowe¹, Enrique Cepero¹ & Gerard Evan²

¹Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

²Cancer Research Institute, University of California at San Francisco, San Francisco, California 94143, USA (e-mail: lowe@cshl.edu)

Mutations that drive uncontrolled cell-cycle progression are requisite events in tumorigenesis. But evolution has installed in the proliferative programmes of mammalian cells a variety of innate tumour-suppressive mechanisms that trigger apoptosis or senescence, should proliferation become aberrant. These contingent processes rely on a series of sensors and transducers that act in a coordinated network to target the machinery responsible for apoptosis and cell-cycle arrest at different points. Although oncogenic mutations that disable such networks can have profound and varied effects on tumour evolution, they may leave intact latent tumour-suppressive potential that can be harnessed therapeutically.

Cancers arise by an evolutionary process as somatic cells mutate and escape the restraints that normally rein in their untoward expansion. Suppressing the emergence of such autonomous cells is an evolutionary imperative of metazoans, particularly in large, long-lived organisms where cells in regenerative tissues retain the potential for neoplastic havoc throughout life. Consequently, multiple mechanisms have arisen to forestall uncontrolled cell division. Some of these are devices within the cell, such as those that limit cell-cycle progression, whereas others are social signals that prompt a cell to remain within its supportive microenvironment. In combination, these tumour-suppressing mechanisms are remarkably effective; on average, cancers arise less than once in a human lifetime, despite trillions of potential target cells, each harbouring hundreds of susceptible cancer-causing genes, all subject to a significant mutation rate. Yet more remarkable is the fact that our tumour-defence systems can discriminate between neoplastic (abnormally growing) and normal cellular states and efficiently quell the former without suppressing the latter.

Insight into the mechanisms that constrain neoplastic progression has come from the realization that many, perhaps all, networks that drive cell proliferation harbour intrinsic growth-suppressive properties. Such innate inhibitory functions obscure any immediate selective advantage that mutations in such pathways might otherwise confer. Because no single pathway confers a net growth advantage, any proto-cancer cell acquiring any single oncogenic mutation is effectively trapped in an evolutionary cul-de-sac. By contrast in normal cells, coordinated extracellular cues activate multiple pathways in concert. In this way, the inherent growth-suppressive activity of each pathway is gated by another, thereby unlocking the cell's proliferative potential (Fig. 1). The nature of the coupling of growth-inhibitory programmes to proliferative networks, and its implications for understanding the evolution and treatment of cancers, are the focus of this review.

Oncogene-induced apoptosis

Cell proliferation and cell death are such diametrically opposed cellular fates that the discovery that the two are linked and interdependent processes was a great surprise^{1,2}. There is little mechanistic overlap between the machineries driving proliferation and apoptosis. Rather, the two processes are coupled at various levels through the individual molecular players responsible for orchestrating cell expansion. Importantly, the same players are often targets for oncogenic mutations, and in many instances, mutations that drive proliferation cooperate with those that uncouple proliferation

from apoptosis during transformation and tumorigenesis^{2,3}. But, although the phenomenon of oncogene-induced apoptosis is now generally accepted as an innate tumour-suppressive mechanism, we have only recently begun to glimpse the diversity and complexity of mechanisms by which oncogenic lesions engage the cell suicide machinery.

At least two distinct general programmes trigger apoptosis, each regulated at many levels (Fig. 2). The 'intrinsic' pathway is the primary death programme responsive to the signals of survival factors, cell stress and injury⁴⁻⁶. The central conduit of this pathway is the mitochondrion, the inter-membrane space of which sequesters a variety of pro-apoptotic effectors that, when released, trigger cellular demise. Mitochondrial permeability is, in turn, determined by the balance between the pro-apoptotic Bax/Bak proteins and their anti-apoptotic Bcl2/BclXL cousins. The activity of these proteins are positively or negatively regulated by the various BH3-only members (Bcl2 family members that contain a single Bcl2 homology-3 domain), each acting as the terminal effector of distinct signalling pathways. According to this simple model, apoptosis occurs when the protective Bcl2/BclXL buffer is breached by the sum of all the active BH3-only proteins, resulting in the net dominance of the pro-apoptotic Bax/Bak proteins, which then permeabilize the mitochondria to release pro-apoptotic factors. One such factor, cytochrome c, acts together with

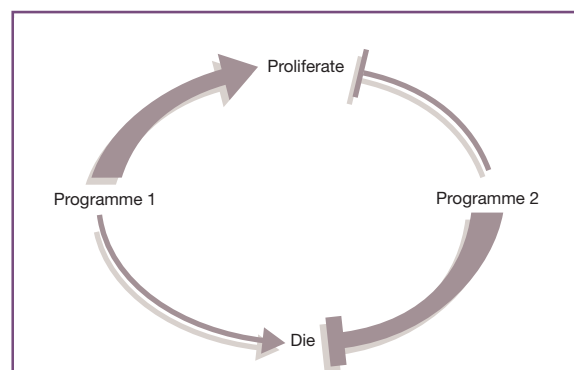


Figure 1 Example of an obligate combinatorial signalling network. Programme 1 drives proliferation and apoptosis, and Programme 2 blocks both. For each cell fate, dominant components are shown as thick lines. Concerted activation of both programmes together leads to cell expansion because Programme 1 overcomes the growth inhibition of Programme 2, and Programme 2 overcomes the lethality of Programme 1. However, activation of either programme on its own triggers cell-death (Programme 1) or senescence (Programme 2).

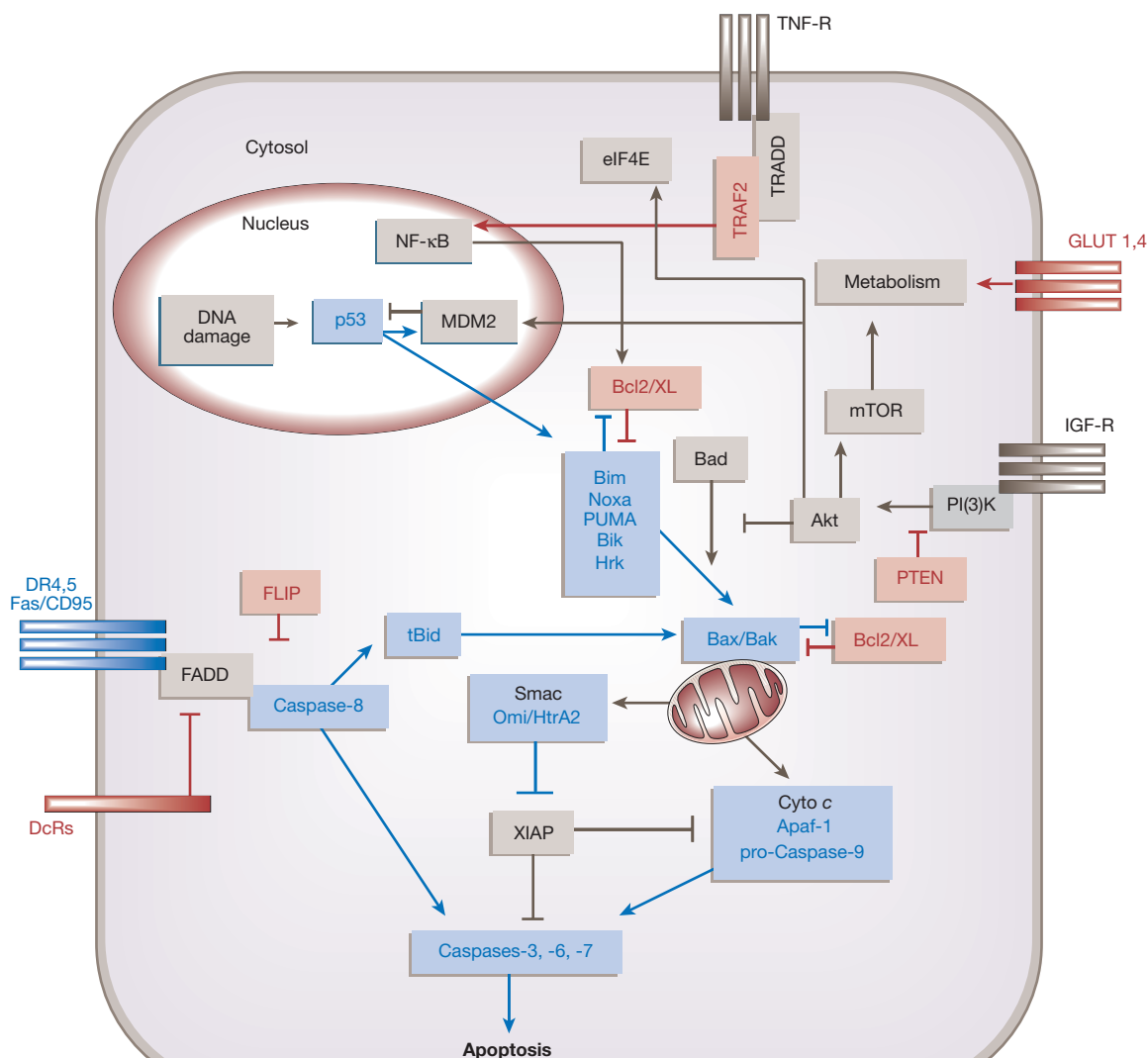


Figure 2 Oncogenic signalling targets many levels of the apoptotic machinery. Shown are key components of the extrinsic and intrinsic apoptotic programmes, as well as some key regulators. Such a network organization allows the cell to sense many aspects of the intracellular and extracellular milieu, yet ensures that cell death

proceeds efficiently once activated. Excessive oncogenic signalling is coupled to apoptosis by a complex mechanism that targets key control points in the pathways. Components highlighted in red can be downregulated by pro-apoptotic oncogenes, whereas components highlighted in blue are often upregulated.

the cell-death adaptor Apaf-1 to trigger the activation of caspase-9, a cysteine protease that initiates a downstream proteolytic cascade that also involves caspase-3 and caspase-7. Once activated, caspases cleave proteins important for cell and genome integrity, orchestrating the orderly death and engulfment of the cell. Regulation of the intrinsic cell-death pathway occurs at many levels, including transcriptional and post-transcriptional regulation of the Bcl2/BH3-only family members, and expression of death-effector components and a class of caspase inhibitors known as ‘inhibitors of apoptosis’ (IAPs).

The ‘extrinsic’ cell-death pathway is activated through ligation of cell-surface ‘death receptors’, such as Fas/CD95, TNFR (tumour necrosis factor receptor) and DR-5, with their respective cognate ligands FasL, TNF α and TRAIL (ref. 7). Once ligated, these receptors form the ‘death-inducing signalling complex’ (DISC), which activates the apical caspase-8. In some cell types, this alone is sufficient to trigger the downstream caspase cascade and consequent apoptosis. In other cells, however, death-receptor-induced apoptosis also requires recruitment of the mitochondrial pathway through caspase-8-mediated activation

of the BH3-only protein Bid (refs 7–9). The extrinsic pathway is subject to modulation by decoy receptors, which bind ligand but are defective in signalling, and by intracellular molecules such as FLIP that compete with caspase-8 for binding to the DISC (ref. 7). In addition, IAPs modulate the activity of both apical and effector caspases in the pathway and, in cells where the intrinsic mitochondrial pathway is co-opted, so do Bcl2 family proteins^{4,8,9}. Remarkably, signals that initiate cell division (mitogenic signals) can interface with the intrinsic and extrinsic programmes at several points.

p53 is a master regulator

p53 is a transcription factor that establishes programmes for apoptosis, senescence, and repair in response to a variety of cellular stresses, including DNA damage, hypoxia, and nutrient deprivation^{3,10}. Known transcriptional targets for p53 in promoting apoptosis include various pro-apoptotic Bcl2 members, including *puma*, *noxa*, *bid* and *bax* (ref. 3), as well as components of death-receptor signalling (for example, DR5, Fas/CD95), the apoptotic-effector machinery

(for example, caspase-6, Apaf-1, PIDD) and others with less well-defined roles (for example, PERP, PML, p53AIP)^{3,10}. Additionally, p53 might directly facilitate cytochrome *c* release¹¹.

p53 is also induced by many oncogenes, including E1A, Myc and E2F (refs 3, 10). Moreover, p53 inactivation severely compromises oncogene-induced apoptosis in many instances. Consistent with this role in coupling proliferation to cell death, inactivation of p53 potentially cooperates with diverse oncogenes to promote transformation *in vitro* and tumorigenesis *in vivo*. For example, p53 inactivation relieves the requirement for E1B in adenovirus transformation of rodent fibroblasts¹², and dramatically potentiates the abilities of Myc, E2F and forms of T antigen that do not bind p53, to promote tumorigenesis in transgenic mouse models^{13–15}. Moreover, studies in mice indicate that selective disruption of the apoptotic machinery downstream of p53 can substitute for p53 loss in promoting tumorigenesis. For example, inactivation of the *bax* or *puma* genes promotes tumorigenesis, despite the presence of wild-type p53, and gene expression of *bcl2* or *bclXL* cooperates with Myc as effectively as p53 loss^{16–19}. Such studies demonstrate that apoptosis is a significant component by which p53 suppresses tumorigenesis.

An especially important mediator of oncogene-dependent activation of p53 is the tumour suppressor ARF (refs 20, 21). Thus, the ability of Myc and E1A to activate p53 is severely compromised in ARF-null cells, which consequently show marked resistance to apoptosis following withdrawal of growth factors^{22,23}. By contrast, ARF is not required for the p53-dependent response to DNA damage²⁴, although it might contribute to a more robust response to DNA damage in oncogene-expressing cells through its stabilization of p53 (refs 23, 25). *In vivo* studies confirm the importance of ARF for oncogene signalling to p53. Disruption of ARF in mice dramatically accelerates Myc-induced lymphomas and carcinomas in a manner broadly comparable to p53 inactivation^{13,26,27}. Deregulated expression of Bmi-1, a polycomb group protein that acts as a negative regulator of the *INK4a/ARF* genetic locus, similarly accelerates Myc-induced tumours²⁶.

Despite its importance, ARF is not the sole conduit through which oncogenes signal to p53. Indeed, in some mouse models of tumorigenesis, ARF inactivation does not appreciably accelerate oncogene-initiated tumorigenesis, even though loss of p53 does^{28,29}. Some evidence suggests that oncogenes can induce genotoxic stress directly and thereby activate p53. Consistent with this, studies also suggest that lesions in DNA-damage repair and response machinery can compromise oncogene-induced apoptosis^{30,31}. However, the relevance of oncogene-induced DNA damage, and whether disruption of the DNA-damage response eliminates oncogene surveillance mechanisms *in vivo*, remains unclear³². Perhaps the machinery that senses DNA damage also mediates responses to non-genotoxic signals that might accompany increased proliferation or transformation, such as an increased nuclear/cytoplasmic ratio³³. Alternatively, given the well-known synergy between DNA damage and pro-apoptotic oncogenes in promoting cell death³, ablating the DNA-damage component might confer protection from apoptosis without it being involved directly in the relationship between the activated oncogene and p53.

p53-independent mechanisms of apoptosis

Although p53 has gained legendary status as our principal defender against malignancy, there are other parallel networks connecting proliferation and apoptosis. The p53 gene itself is a member of a family that includes p63 and p73, both of which encode proteins implicated in apoptosis and several other processes⁴. Disruption of p63 or p73, either alone or in combination, ameliorates apoptosis in cultured fibroblasts³⁴, and both can induce p53 transcriptional targets and apoptosis when overexpressed^{35,36}. Moreover, p73 is a direct transcriptional target of E2F and Myc, and both p63 and p73 can act to redirect p53 to the promoters of pro-apoptotic genes³⁴, although such mechanisms are not universal³⁷. In addition, p73 can promote

apoptosis in p53-deficient cells, a property that can be blocked by a specific cadre of 'gain-of-function' p53 mutants that are able to associate physically with p73 (refs 38–40). Nonetheless, mice heterozygous for either p63 or p73 are not overtly tumour-prone^{35,36}, so the exact extent of the contribution of the p53 siblings to tumour suppression *in vivo* remains uncertain.

Oncogenes can also target various components of the cell-death machinery independently of p53. Thus, E1A suppression of FLIP sensitizes cells to death-receptor-induced apoptosis⁴¹, and Myc sensitizes cells to death-receptor signalling by recruiting the mitochondrial pathway⁴². Additionally, Myc, E2F and E1A have pleiotropic effects on the expression of pro- and anti-apoptotic members of the Bcl2 family. For example, Myc represses expression of *bcl2* and *bclXL*, whereas E1A and E2F suppress another Bcl2-family gene *mcl-1* (refs 43, 44). Myc and E2F also induce expression of several BH3-only killer proteins, including Bim (refs 45, 46). Finally, E2F can induce several downstream effectors of the apoptotic machinery, including various caspases⁴⁷.

The relative importance of p53-independent versus p53-dependent apoptotic mechanisms in suppressing tumorigenesis remains unclear. Experimental overexpression of Bcl2 can relieve selection against loss of p53 in the Eμ-Myc mouse model of lymphoma¹⁸ (where the myc oncogene is expressed from an immunoglobulin enhancer), whereas inactivation of even a single allele of *bim*, a Myc target, dramatically accelerates Myc-induced lymphomagenesis in the same model⁴⁵. Nonetheless, it seems likely that the relative contributions of p53-dependent and p53-independent apoptotic pathways will vary depending on tumour type, and on the nature and sequence of oncogenic mutations within any specific cancer.

Overlapping mechanisms of oncogene-induced cell death

If oncogenic lesions engage a variety of effector molecules that modulate cell proliferation in diverse ways, then how can each be coupled to the same core death programme? The simplest possibility is that pro-apoptotic oncogenes act at different points in a single linear pathway that is coupled to apoptosis through some downstream node. The mechanisms by which the Myc and E2F oncogenes promote apoptosis illustrate this point. Myc activates E2F, and there are consensus Myc-binding sites in at least one E2F promoter⁴⁸. Moreover, in cultured mouse embryonic fibroblasts (MEFs), Myc-induced apoptosis can be dependent on E2F1 (refs 49, 50). Accordingly, both deregulated Myc expression and inactivation of the tumour-suppressing retinoblastoma (Rb) protein exert some of their apoptotic action through common downstream E2F effectors.

Still, there are clear differences in the apoptotic modes of action of Myc and E2F. For example, both *in vitro* and *in vivo* studies indicate that ARF is more important for apoptosis induced by Myc than for that induced by ectopic E2F expression or Rb inactivation^{29,51}. Consistent with this, Myc can induce ARF through E2F1-independent mechanisms⁵². Furthermore, E2F1, but not Myc, augments apoptosis following cytosolic injection of holocytochrome *c*, indicating that E2F directly influences components of the apoptotic effector machinery downstream of the mitochondrial switch⁴⁷. Indeed, E2F1 directly controls the expression of certain caspases⁴⁷, an activity that Myc does not share (Z. Nahle and S. W. L., unpublished work).

In reality, dissecting the precise interrelationship between Myc and E2F1 in apoptosis signalling is complicated by the multiplicity of E2F proteins, each of which can induce and compensate partially for the others, at least when overexpressed^{53,54}. Indeed, that Myc and E2F normally act in a highly integrated signalling network makes it difficult, even in principle, to assign individual contributions to each. Probably, Myc and E2F promote apoptosis by targeting multiple processes (some that converge on common targets and others that are distinct), that then act collectively to engage the apoptotic programme. It appears that the cell-proliferative and cell-death machineries are not coupled through a single conduit but that evolution has employed a variety of redundant mechanisms to link the two.

Coupling proliferation to senescence

Apoptosis is not the only anti-proliferative response coupled to oncogenic signalling. Activated oncogenes can also trigger cellular senescence^{55–57}, a state characterized by permanent cell-cycle arrest and specific changes in morphology and gene expression that distinguish the process from quiescence (reversible cell-cycle arrest)^{58,59}. Whereas 'replicative' senescence is triggered by the erosion of telomeres during cell divisions, a similar phenotype can occur in 'young' cells in response to oncogenes, DNA damage or oxidative stress^{58,59}. Consistent with their roles in mediating cell-cycle checkpoints and tumour suppression, both Rb and p53 tumour suppressors are key regulators of the senescence programme.

Oncogenic Ras promotes cellular senescence in non-immortal human and rodent cells in a manner that depends on one or both products of the *INK4a/ARF* locus, which encodes the tumour suppressor proteins p16 and ARF (Fig. 3)^{20,21}. The mitogen-activated protein kinase (MAPK) signalling cascade appears to be the principal Ras-effector pathway responsible for cellular senescence by inducing p16 and/or ARF, and ultimately by activating Rb and p53, respectively^{20,59}. p53 and Rb then promote senescence by controlling a number of effectors, including p21CIP1/WAF1, PML, and various chromatin-modifying factors that produce a repressive state that buffers proliferative genes from mitogenic signalling^{60–64}. The respective contributions of Rb and p53 to senescence are apparently cell-type dependent: thus, MEFs depend primarily on the ARF–p53 axis, whereas human fibroblasts and some rodent haematopoietic cells also rely on p16–Rb functions²⁰.

Escape from oncogene-induced senescence is a prerequisite for the transformation of cells that probably explains the oncogenic cooperation between Ras and so-called 'immortalizing' oncogenes *in vitro*. Thus, in mouse embryonic fibroblasts or dermal keratinocytes, disruption of either ARF or p53 abrogates Ras-induced cytostasis and permits oncogenic transformation^{20,21}. In human cells, the situation is more complex, often requiring additional oncogenic lesions to thwart senescence; for example, *INK4a* loss^{20,21}. High Ras levels are frequently observed in tumour cells and are probably required for malignant conversion⁶⁵. Cancers must therefore acquire cooperating lesions that uncouple mitogenic Ras signalling from senescence. Such secondary lesions that thwart senescence are likely to be required for tumour maintenance, as suggested by the observation that suppression of the p53-inactivating E6 oncoprotein rapidly triggers senescence in human cervical carcinoma cells⁶⁶.

In general terms, both senescence and apoptosis seem to serve the same ends in tumour suppression. Both represent an irrevocable growth-inhibitory cellular response to oncogenic stress that acts as a potent barrier to the further evolution of any pre-neoplastic cell. Indeed, many of the signals that promote apoptosis in one cell type induce senescence in others. For example, both E2F and Myc can be either pro-apoptotic or pro-senescent depending on the cell type, the levels to which they are expressed, and the extent of other pro-apoptotic and growth signals received by the cell^{55,67}. It is plausible that both programmes are induced by the same generic processes and have been structured by evolution to serve as backups for each other.

Is senescence relevant?

Although it is generally accepted that oncogene-induced apoptosis is a bona fide tumour-suppressor mechanism, the role of oncogene-triggered senescence is more contentious because the programme has not been observed definitively *in vivo*. Even *in vitro*, oncogenic Ras does not always trigger senescence in primary cells⁶⁸. This is most notable when it is expressed from its endogenous locus^{69,70}, raising the troubling possibility that the whole phenomenon of Ras-induced senescence is an artefact of overexpression *in vitro*. Such a possibility has devastating ramifications, because most of our current understanding of genetic interactions in cancer depends on studies involving Ras overexpression. Unfortunately, defining any role for senescence in tumour suppression *in vivo* is complicated by extreme difficulty in

identifying senescence *in vivo*, and our relatively rudimentary understanding of the mechanisms that regulate it.

Studies in mouse models provide circumstantial evidence that senescence acts to counter tumorigenesis induced by mitogenic mutations. In chemically induced skin carcinogenesis in mice, the initiating carcinogen induces mutations in the endogenous *H-ras* gene in multiple target cells. However, progression of such incipient proto-tumour cells into malignant tumours requires obligate secondary mutations in the *p53*, *p16INK4a*, *ARF* or *p21CIP1/WAF1* genes—precisely those that mediate Ras-induced growth arrest in cultured dermal keratinocytes^{71,72}. Likewise, enforced E2F expression in the mouse pituitary gland initially promotes proliferation and tissue expansion that then stalls because of a progressively increasing insensitivity of the affected cells to further mitogenic stimulation (K. Helin, personal communication). The non-dividing tissue displays significant upregulation of p16 and other markers of senescence—offering strong evidence that a senescence programme suppresses aberrant proliferation *in vivo*.

Both apoptosis and senescence involve integrating diverse extracellular and intracellular influences into a binary live/die or go/stop cellular decision. For example, we know that Myc-induced apoptosis is a contingent phenomenon that is potentially inhibited by survival factors and greatly exacerbated by additional insults with pro-apoptotic signals. In effect, Myc activation contributes only one component to the net pro-apoptotic load of any individual cell: whether that is enough to breach the apoptotic firing threshold depends on a host of contributing factors including level of Myc expression, cell type, location and availability of trophic survival signals, and differentiation and stress status. Such contingency is clearly observed in studies of transgenic mice that show that, when activated *in vivo*, Myc is a powerful destroyer of certain cell types but not others^{19,73}. By analogy to apoptosis, therefore, we might expect senescence to be dramatically influenced by the cellular micro-environment. Consequently, Ras might induce senescence only in certain cell types and, even then, perhaps only in combination with other simultaneous insults, such as DNA damage or growth-factor deprivation. Such 'contingent' senescence would not be readily apparent in conventional transgenic studies, but it could explain why oncogenic Ras, similar to Myc, is only capable of directly inducing tissue expansion in a subset of tissues⁷⁰.

Crossing thresholds

The decisions whether to live or die, to proliferate or arrest are choices a cell must make in the face of many disparate influences. Furthermore, once a threshold for firing such programmes has been breached, they necessarily run to completion. During oncogene-induced apoptosis, the threshold is probably crossed when the pro-apoptotic influences far outweigh the anti-apoptotic buffer (Fig. 3). The ability of BH3-only proteins to integrate apoptotic signals offers an explanation of why diverse stimuli, such as DNA damage, death-receptor signals and activated oncogenes show synergy^{2,4}. One consequence of such signal integration is that it is neither possible nor meaningful to attribute the ultimate outcome to any one signal, because elimination of any single component might be sufficient to drop the system below the firing threshold. When operating close to its firing thresholds, the relative contributions of individual components to a particular biological process are not additive.

The lessons of thresholds in the control of apoptosis are important. Just because deletion of a specific gene causes a 90% reduction in apoptosis does not mean that all the other pro-apoptotic influences together account for the remaining 10% of cell death. Accordingly, a mutation in any one of the downstream pathways by which oncogenes promote apoptosis might be sufficient to suppress a significant degree of cell death and so confer a significant growth advantage. Such a scenario might explain how deletion of *bim* accelerates Myc-induced lymphomagenesis or compensates for spontaneous *p53*

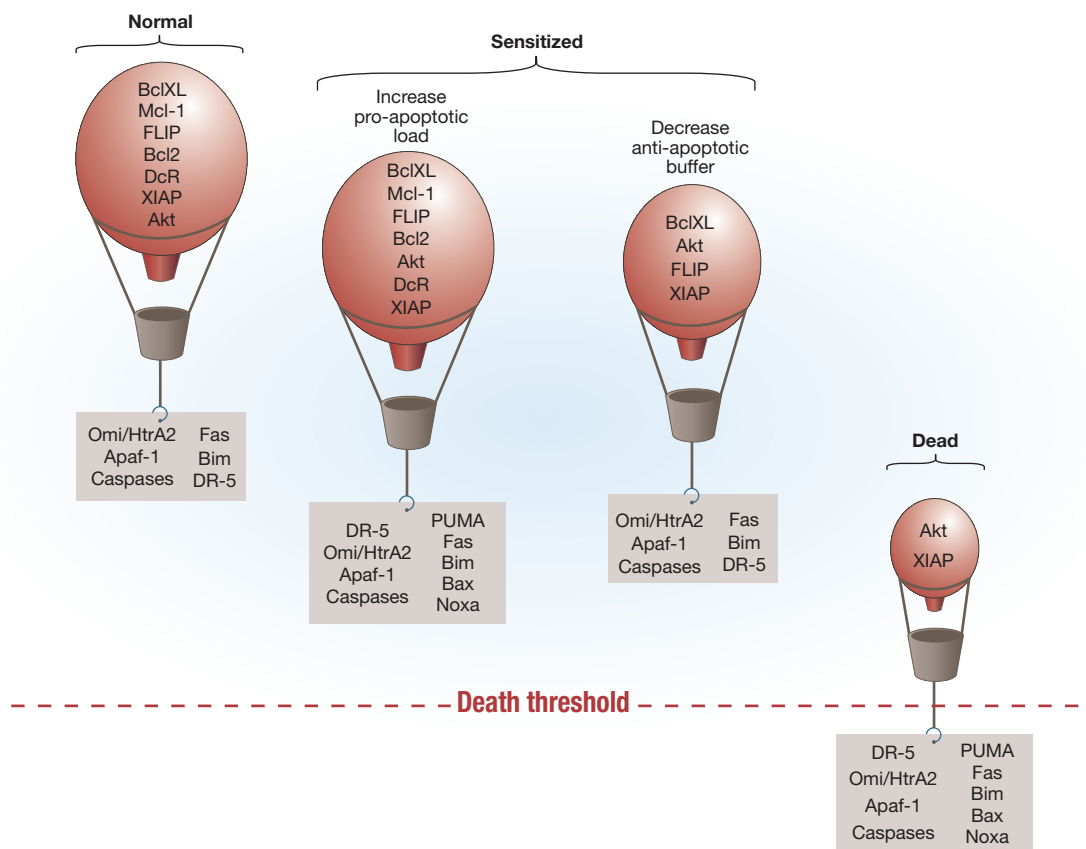


Figure 3 Crossing the apoptotic threshold. Apoptosis is tightly controlled by the ability of the cell to integrate many pro- and anti-apoptotic signals into a binary life/death decision. In one model, apoptosis occurs when the pro-apoptotic load of the cell exceeds its anti-apoptotic buffering capacity, breaching a threshold that then triggers an effector programme capable of running to completion.

Perhaps the most important feature of a threshold model is that it is impossible to attribute the ultimate outcome to one specific signal. Elimination of any individual component could be sufficient to shift the entire system below the firing threshold. Where thresholds operate, the contributions of individual components are not additive.

mutations, even though *bim* is not induced by p53 (ref. 45). Loss of the Bim protein presumably drops the cell below its apoptotic threshold and allows cell survival in the presence of wild-type p53. Importantly, to bring that cell back to its apoptotic firing threshold it might not be necessary to correct that specific lesion or modulate that specific pathway — adding to the general apoptotic load through other pathways could be equally effective.

Sensing aberrant proliferation

Much of the above discussion has focused on how oncogenic signalling interfaces with the cell death or senescence machineries. However, apoptosis and senescence are not the inevitable outcome of normal cell division, but are mostly confined to aberrantly proliferating cells. The implication is that specific molecular sensors determine whether proliferation is aberrant, implying that concrete criteria must distinguish normal and abnormal cell proliferation. Understanding the nature of such criteria and how they are sensed would provide insight into both the selectivity of tumour suppression and the generic self-organizing rules that craft and maintain normal somatic tissues.

At least two general mechanisms have been identified by which cells and their adjacent tissues might 'sense' which cells are cancerous. One depends upon the obligatory social dependency that somatic cells possess for specific microenvironmental trophic signals, effectively using the orthotopic disposition of cells in tissues as cues of their normalcy. The other appears to involve some kind of internal registry of normal and abnormal proliferative signal strengths, triggering only in response to the latter. Both mechanisms

appear to work in concert to limit the transforming potential of mitogenic oncogenes.

Microenvironmental signals

Somatic cells are thought to be continuously dependent upon their neighbours and local microenvironment to provide them with trophic signals that quell their innate suicidal tendencies⁷⁴. One way that activated oncogenes trip the tumour-suppressive failsafe is by super-activating apoptotic and senescence tumour-suppressor programmes, which then overwhelm the limited social buffering capacity of local trophic factors. In addition, oncogene-induced cell expansion forces cells into inhospitable trophic compartments. Consistent with this, cells expressing mitogenic oncogenes such as Myc, E1A and E2F are peculiarly susceptible to induction of apoptosis upon withdrawal of survival factors, such as the insulin-like growth factors I and II (IGF-1, IGF-II) in fibroblasts, or interleukin-3 in myeloid cells^{75,76}. In epithelial cells, survival signals are also derived from the association with the extracellular matrix. This is evident in the basal epidermis and intestinal epithelium where obligate survival signals are provided by the basal lamina⁷⁷.

Many survival factors prevent apoptosis by triggering receptor tyrosine kinases that ultimately signal through Ras and the phosphatidylinositol-3-OH kinase (PI(3)K) signalling cascade⁷⁸. A key mediator of PI(3)K signalling is the Akt/PKB kinase, which phosphorylates multiple effectors leading to pleiotropic changes in proliferation, metabolism, cell growth and survival. Akt promotes survival by coordinating programmes that directly inhibit apoptotic effectors,

suppress transcription of pro-apoptotic genes, and modulate the translation of cell-death regulatory messenger RNAs⁷⁸. Additionally, Akt survival signalling is potentiated by its effects on cellular bioenergetics⁷⁹, and its modulation of the mTOR pathway, which controls the cell response to nutrients⁷⁹. Some cytokines also trigger PI(3)K-independent activation of STATs and NF- κ B, transcription factors that promote cell survival by modulating the transcription of the Bcl2-related proteins and other anti-apoptotic genes⁸⁰.

Because limited trophic support restricts tissue expansion, it is not surprising that mutations that constitutively activate survival-signalling pathways contribute to the neoplastic genotype. Thus, elevated signalling through the IGF pathway occurs in many tumour types⁸¹, and IGF-II availability is required for progression of oncogene-induced insulinomas in mice⁸². Similarly, genetic lesions that activate various elements of the PI(3)K pathway dramatically cooperate with Myc during cancer development^{83,84}. Such mutations ameliorate the dependency of incipient tumour cells for their normal somatic compartments as well as acting as generic suppressors of apoptosis that render cells less susceptible to stress and micro-environmental changes⁸⁵.

The oncogene checkpoint

Although social circumstances can greatly influence the expansion of normal and pre-neoplastic cells, cells also harbour pre-set and autonomous sensors for aberrant proliferative signalling²⁰. Such sensors discern elevated or sustained fluxes of mitogenic signalling, much like stress response 'checkpoints', and generally respond through the p53 pathway. One of the most important of these sensors is ARF, which, as described earlier, is transcriptionally upregulated in response to many oncogenes²⁰. ARF is not expressed in normal proliferating tissues, but is rapidly induced in response to aberrant signals such as activated Myc (ref. 86). Thus, ARF expression is buffered against normal mitogenic signalling, becoming active only when some preconfigured signalling threshold is exceeded. This explains why, even though Myc and E2F are activated during the course of cell-cycle progression, ARF is not a cell-cycle regulated gene²⁰.

Factors that control ARF expression provide clues to the nature of this buffering threshold. In normal cells, the ARF promoter is actively suppressed by E2F3b, a variant of E2F3 that acts as a transcriptional repressor⁸⁷. However, in the presence of E1A or elevated E2F1, E2F3b is displaced from the ARF promoter, allowing the binding of activator E2Fs. What signals this transition remains to be determined, but such observations provide the first clear evidence of an absolute difference in ARF regulation in normal cells versus oncogene-expressing cells. Nonetheless, whereas deletion of E2F3 upregulates ARF in cultured fibroblasts, the same does not occur *in vivo*, implying the existence of additional mechanisms insulating ARF during normal mitogenesis. One probable mechanism involves control of the polycomb group protein Bmi-1 — a chromatin remodelling factor that is an established repressor of the *INK4a/ARF* locus²⁶. Perhaps sustained oncogenic signalling suppresses Bmi-1 function, producing a more open chromatin structure that enables activation of the ARF promoter by mitogenic transcription factors such as E2F1.

Deconstructing the network

From the above it is clear that oncogenic mutations can inhibit proliferation through a variety of mechanisms. Although it is possible that each acts to trigger apoptosis or cellular senescence under a specific set of circumstances or in certain cell types, it seems unlikely that evolution would have incorporated so many disparate means to achieve the same end. Instead, a more likely explanation for this mechanistic diversity is that each pathway or signal transducer acts as part of a complex network that coordinates the processes of apoptosis and senescence by targeting each programme at multiple levels. Through this organization, the cell ensures that the process is not dependent on a single event and proceeds efficiently once engaged.

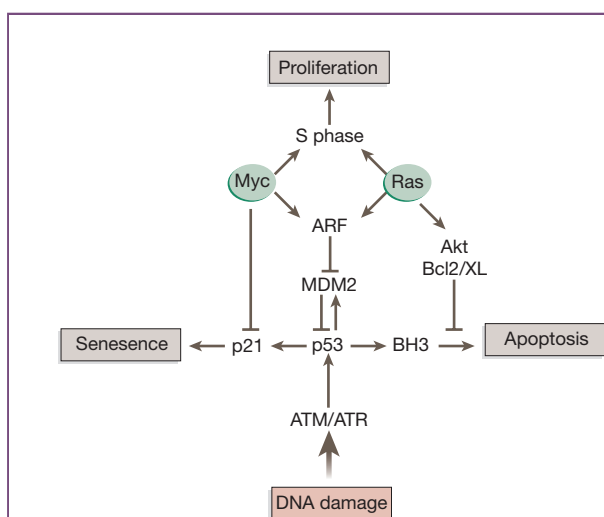


Figure 4 The ARF–p53 circuit in tumour development and therapy. Activation of Myc and Ras can force proliferation or trigger apoptosis or senescence. These oncogenic signals engage the tumour-suppressor network at many points, including through the ARF–p53 circuit shown here. Which components contribute most to tumour suppression depends on context. For example, Myc activates p53 to promote apoptosis while interfering with its ability to induce senescence. Conversely, Ras activates p53 to promote senescence while suppressing apoptosis. This simplified view helps explain why, despite the potential of p53 to control several processes, apoptosis is primarily responsible for p53-mediated tumour suppression in the presence of Myc, and why mutations that disable apoptosis (for example, Bcl2 overexpression) cooperate more effectively with Myc than Ras. As another example, DNA damage and oncogene signalling engage the tumour-suppressor network at different points and, as such, DNA-damage signalling relies more on p53 than on ARF to elicit an anti-proliferative response. Such a model explains why loss of ARF or p53 confer similar advantages during Myc-induced tumorigenesis but not following treatment with DNA-damaging drugs. Here, drug resistance is an unselected trait conferred by p53 mutations that provides a unique advantage as the tumour encounters a new environment (for example, chemotherapy).

By revisiting some of the mechanisms whereby Myc promotes apoptosis, it is possible to envision how an oncogene-triggered tumour-suppressor network might act to coordinate a cell-death programme effectively. By greatly increasing the ratio of pro- to anti-apoptotic Bcl2 proteins, Myc promotes mitochondrial permeabilization and release of cytochrome *c*. Through indirectly activating p53 or E2F, Myc induces Apaf-1, caspases, and the IAP inhibitor Omi/Htr2 (refs 47, 88), and consequently increases the efficiency with which cytochrome *c*, when released, triggers the caspase cascade. Through p53-dependent increases in PTEN, Myc might short-circuit survival signalling, thereby reducing the cell's ability to buffer pro-apoptotic signals. And, by indirectly increasing death receptors and decreasing their antagonists, Myc sensitizes the cell to the actions of death-inducing ligands in the microenvironment. Finally, by upregulating p73, Myc introduces redundancy in the p53-dependent programme, reinforcing many of the processes described above.

Sorting out how individual components of such a complex and multifarious network contribute to the output of each programme networks is a major challenge because, by definition, analysing individual components in isolation cannot provide a complete picture of network dynamics. Biological networks are characterized by multiple feed-forward, feedback, and cross-talk characteristics that compensate for perturbations affecting individual components and lend them great robustness. Consequently, the phenotype caused by disrupting a specific protein might reflect not its normal function but, rather, the net difference between its activity and an opposing compensatory signal. This probably explains the failure of apparently important

genes to produce profound phenotypes when deleted in mice, and it will complicate efforts to assign specific roles to certain caspases or IAPs in apoptosis, or Rb-family members in promoting senescence^{89–91}.

The situation is more complex still given the pivotal importance of cell type and cellular microenvironment in determining the net impact of oncogenic mutations. Cell-type-specific levels of endogenous pro- and anti-apoptotic effectors, together with microenvironmental factors and other oncogenic events, all influence the signalling flux through pathways that contribute to cell proliferation and viability. Consequently, the neoplastic impact of any oncogenic mutation is likely to lead to dramatically different outcomes depending on context. For example, acute activation of Myc in pancreatic β -cells leads to rapid β -cell involution and diabetes¹⁹. By contrast, activation of Myc in skin triggers proliferation without cell death, probably because of an abundance of local survival factors, resulting in rapid development of papillomatous hyperplasias⁷³. In the latter circumstance, anti-apoptotic lesions such as that caused by the loss of p53 exert a selective advantage only when the neoplastic cell moves beyond its normal trophic environment into the dermis¹⁹. By the same token, mutations that inhibit death-receptor signalling would only enhance viability in environments where death ligands are present.

The multi-functionality of individual signalling molecules adds a final tier of complexity to signalling networks and how they drive tumour evolution. For example, although Myc and Ras both engage the ARF/p53 pathway, they also instigate distinct 'collateral signals' that elicit different outcomes following p53 activation. Although Myc induces apoptosis, it also overrides cell-cycle arrest⁹²; thus, subsequent immortalizing mutations provide no further selective advantage. Conversely, Ras promotes senescence yet attenuates apoptosis, rendering subsequent anti-apoptotic mutations mostly redundant¹⁷ (Fig. 4). Thus, both the context and sequence of mutations profoundly influence the trajectory of tumour evolution context, and so determine which lesions end up as most critical for maintenance of the end-stage tumour.

From compliance to autonomy in tumour evolution

One peculiar consequence of all interlocking networks is that any single mutation in such a network can engender adventitious traits that, although having no immediate impact, might confer selective advantages (or disadvantages) later. A pertinent example of such pre-adaptation, or 'exaptation', relates to the impact of p53 mutations on tumours arising in E μ -Myc transgenic mice^{13,18}. In this model, directed mutations that inactivate p53 or PUMA, or that overexpress Bcl2, dramatically accelerate tumorigenesis. However, although tumours that arise in each case are phenotypically similar and display broadly equivalent apoptotic defects, only those with mutant p53 display defects in DNA-damage checkpoints and gross aneuploidy^{13,18}. Moreover, E μ -Myc lymphomas lacking p53 progress to a lethal stage more rapidly than those overexpressing Bcl2, presumably because loss of p53 confers a selective advantage under conditions of checkpoint activation or genomic damage^{13,17,18}. Thus, although disruption of Myc's apoptotic function is directly selected during lymphomagenesis, the mechanism by which it occurs influences the future evolution of the tumour as it encounters new stresses or environments. Importantly, what is crucial and what is an evolutionary byproduct will depend on context: different rules are likely to apply to selection against p53 action in, say, suppressing Ras-induced tumorigenesis¹⁷. Such considerations have important implications for how tumours respond to therapy, both initially and evolutionarily.

Evolving towards drug resistance

The major limitation to conventional cancer therapy is drug resistance, either because the initial tumour fails to respond to therapy or because it acquires resistance during relapse. Most conventional chemotherapeutic agents damage cellular components, and it was long assumed that this damage was directly responsible for the anti-tumour effect. However, damage induced by chemotherapeutic

drugs is not invariably lethal but instead actively triggers damage responses (often apoptosis or senescence), and it is these responses that determine the eventual fate of the cell⁹³. Ironically, classical cancer therapies unwittingly exploited the very same innate tumour-suppressor networks that suppress aberrant cell proliferation.

The fact that oncogenes and conventional cancer drugs both co-opt the same networks means that mutations that uncouple proliferation from apoptosis and senescence can disable drug responses. In the E μ -Myc lymphoma model, inactivation of p53 confers an immediate advantage to the tumour by suppressing cell death, and predisposes the tumour to a poor response to chemotherapy¹³. Here, drug resistance is not the directly selected trait, but another example of exaptation. Similar forces could explain the innate drug resistance of some of the more aggressive primary tumours⁹³, as well as the link between p53 loss and other tumour-promoting mutations resulting in drug resistance in certain human cancers⁹³.

Although different oncogenes might cause similar phenotypes, drug responses will differ depending on the way these oncogenes affect the cellular signalling network and on network components targeted by different cancer drugs (Fig. 4). For example, mutations in either ARF or p53 are frequent in E μ -Myc-dependent lymphomagenesis, and targeted disruption of either gene yields accelerated and highly aggressive malignancies. But despite such overt similarities, ARF-null tumours undergo massive apoptosis and are frequently cured following treatment with cyclophosphamide, a DNA-damaging drug, whereas p53-null tumours respond poorly²⁵. This can be explained by the way Myc and DNA-damage signals engage the tumour-suppressor network; whereas Myc signalling depends heavily on ARF, DNA damage does not^{22,24}. However, p53 is important for the DNA-damage response. Additionally, the genomic instability conferred by p53 loss, less pronounced in ARF-null or Bcl2-overexpressing tumours, might bestow an additional advantage under therapy. Similar principles could contribute to the enormously heterogeneous response to therapy observed in human cancer patients.

Curiously, the very same interrelationship between oncogene signalling and drug action could explain the remarkably selective ability of conventional chemotherapeutic drugs to kill tumour cells. Tumour cells harbour mitogenic lesions that drive their proliferation but also confer a propensity towards apoptosis or senescence. Hence, established tumours reside substantially closer to the threshold at which apoptosis or senescence can be triggered (Fig. 3). By comparison, normal somatic cells, lacking oncogenic mutations and protected by the trophic signals within their orthotopic environments, are far from such thresholds and consequently less susceptible to the cytotoxic and cytostatic effects of therapeutic agents.

In summary, the fact that proliferation is coupled to apoptosis and senescence coerces the evolutionary trajectory of tumours in ways that influence cellular responses to therapy, by promoting drug resistance or, conversely, by increasing the probability that the drug will be effective. By understanding these relationships, the hope is that current cancer therapies can be employed more effectively. A subsidiary question concerns the extent to which similar rules apply to the new targeted therapeutics that target key oncoproteins or their effectors. However, in many instances these novel agents induce apoptosis, raising the possibility that they may act, in part, by hijacking existing tumour-suppressor networks.

Exploiting the Achilles' heel of cancer cells

Re-engaging the disrupted senescence and apoptosis programmes by novel targeted therapeutics in cancer cells offers a compelling general strategy for effective and tumour-cell-specific cancer therapy. However, it will only work if the engines driving apoptosis and senescence persist throughout the lifetime of the tumour. Thus traits selected early on in the neoplastic process might not remain under continuous selection during tumour progression and may even be selected against at later stages. Fortunately, a number of studies using conditional transgenic and knockout mouse models indicate that the

initiating oncogenic lesions remain essential for tumour maintenance^{19,73,94,95}, even when the tumours have evolved to an advanced stage^{2,19}. Indeed, even in situations where additional collaborating mutations appear to be required to sustain tumorigenesis^{96,97}, it seems likely that rational targeting of one or a few pivotal oncogenic lesions would undermine the entire neoplastic edifice.

Another of the remarkable features of such conditional transgenic and knockout mouse models is that dominant oncogenes such as Myc and Ras are capable of driving multiple aspects of advanced tumorigenesis, including angiogenesis and invasion, when unshackled from their inherent apoptosis and senescence programmes. By the same token, the necessity for pre-neoplastic cells to evolve mechanisms to quell their innate predisposition to apoptosis or senescence exposes a critical and exploitable chink in their defences. As well as being addicted to their initiating oncogenic mutations, tumour cells will remain critically dependent on their limited repertoire of anti-apoptotic and anti-senescent mutations; by contrast, normal cells, lacking pro-apoptotic oncogenic lesions and safely ensconced in their stress-free, trophic havens, will not. Consequently, tumour cells appear particularly sensitive to interventions that re-establish pro-apoptotic pathways or disable survival programmes.

Recent *in vivo* studies illustrate that re-engaging apoptotic programmes disabled during tumour evolution can indeed have a profound therapeutic effect⁸⁵. Thus, inhibition of Bcl2, or reactivation of p53, has proven particularly lethal to appropriate tumour types^{98,99,100}. Likewise, rapamycin, an inhibitor of the Akt target mTOR, effectively reverses resistance to conventional chemotherapy in Eμ-Myc lymphomas co-expressing Akt (ref. 85). Importantly, rapamycin works selectively only in tumours where the apoptotic fail-safe has been ablated by Akt — not those where apoptosis has been disengaged through lesions that act in parallel to, or downstream of, mTOR (ref. 85). Such studies intimate that the effective use of similar strategies in human patients will require significant insight into the evolution of each individual's neoplastic disease. Nonetheless, there seems no doubt that harnessing the very mutations that cancer cells need to promote their pathological survival and expansion will be the basis of the therapeutic strategy of the future.

Perspective

Cancer has long been considered to be an endlessly adaptable and profoundly complex disease treatable only with blunt approaches that frequently do as much damage to the patient as to the tumour. Contemporary molecular dissection of tumour cells has confirmed the complexity and subtlety of the signalling networks that drive and maintain tumours, but it has also shown us that tumour cells harbour the seeds of their own potential destruction: the very oncogenic mutations that cancer cells need to drive their relentless and pathological expansion possess the potential to unleash powerful tumour-suppressor programmes such as senescence and apoptosis. Cancers arise when the molecular network connecting proliferation and tumour suppression become uncoupled. Even then, however, the underlying tumour-suppressor programmes remain intact, awaiting only adroit human intervention to reconnect them and herald a new era of effective and tumour-specific therapies. □

doi:10.1038/nature03098

1. Evan, G. I. *et al.* Induction of apoptosis in fibroblasts by c-myc protein. *Cell* **69**, 119–128 (1992).
2. Evan, G. I. & Vousden, K. H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348 (2001).
3. Fridman, J. S. & Lowe, S. W. Control of apoptosis by p53. *Oncogene* **22**, 9030–9040 (2003).
4. Danial, N. N. & Korsmeyer, S. J. Cell death: critical control points. *Cell* **116**, 205–219 (2004).
5. Cory, S., Huang, D. C. & Adams, J. M. The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene* **22**, 8590–8607 (2003).
6. Green, D. R. & Kroemer, G. The pathophysiology of mitochondrial cell death. *Science* **305**, 626–629 (2004).
7. Peter, M. E. & Krammer, P. H. The CD95 (APO-1/Fas) DISC and beyond. *Cell Death Differ.* **10**, 26–35 (2003).
8. Wilkinson, J. C., Cepero, E., Boise, L. H. & Duckett, C. S. Upstream regulatory role for XIAP in receptor-mediated apoptosis. *Mol. Cell Biol.* **24**, 7003–7014 (2004).
9. Scaffidi, C. *et al.* Two CD95 (APO-1/Fas) signalling pathways. *EMBO J.* **17**, 1675–1687 (1998).

10. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
11. Mihara, M. *et al.* p53 has a direct apoptogenic role at the mitochondria. *Mol. Cell* **11**, 577–590 (2003).
12. White, E. Regulation of the cell cycle and apoptosis by the oncogenes of adenovirus. *Oncogene* **20**, 7836–7846 (2001).
13. Schmitt, C. A., McCurrach, M. E., de Stanchina, E., Wallace-Brodeur, R. R. & Lowe, S. W. INK4a/ARF mutations accelerate lymphomagenesis and promote chemoresistance by disabling p53. *Genes Dev.* **13**, 2670–2677 (1999).
14. Symonds, H. *et al.* p53-dependent apoptosis suppresses tumour growth and progression *in vivo*. *Cell* **78**, 703–711 (1994).
15. Pierce, A. M. *et al.* Increased E2F1 activity induces skin tumours in mice heterozygous and nullizygous for p53. *Proc. Natl Acad. Sci. USA* **95**, 8858–8863 (1998).
16. Yin, C., Knudson, C. M., Korsmeyer, S. J. & Van Dyke, T. Bax suppresses tumorigenesis and stimulates apoptosis *in vivo*. *Nature* **385**, 637–640 (1997).
17. Hemann, M. T. *et al.* Suppression of tumorigenesis by the p53 target PUMA. *Proc. Natl Acad. Sci. USA* **101**, 9333–9338 (2004).
18. Schmitt, C. A. *et al.* Dissecting p53 tumour suppressor functions *in vivo*. *Cancer Cell* **1**, 289–298 (2002).
19. Pelengaris, S., Khan, M. & Evan, G. I. Suppression of Myc-induced apoptosis in beta cells exposes multiple oncogenic properties of Myc and triggers carcinogenic progression. *Cell* **109**, 321–334 (2002).
20. Lowe, S. W. & Sherr, C. J. Tumour suppression by Ink4a-Arf: progress and puzzles. *Curr. Opin. Genet. Dev.* **13**, 77–83 (2003).
21. Sherr, C. J. The INK4a/ARF network in tumour suppression. *Nature Rev. Mol. Cell Biol.* **2**, 731–737 (2001).
22. Zindy, F. *et al.* Myc signalling via the ARF tumour suppressor regulates p53-dependent apoptosis and immortalization. *Genes Dev.* **12**, 2424–2433 (1998).
23. de Stanchina, E. *et al.* E1A signalling to p53 involves the p19(ARF) tumour suppressor. *Genes Dev.* **12**, 2434–2442 (1998).
24. Kamijo, T. *et al.* Tumour suppression at the mouse INK4a locus mediated by the alternative reading frame product p19ARF. *Cell* **91**, 649–659 (1997).
25. Schmitt, C. A. *et al.* A senescence programme controlled by p53 and p16INK4a contributes to the outcome of cancer therapy. *Cell* **109**, 335–346 (2002).
26. Jacobs, J. J. *et al.* Bmi-1 collaborates with c-Myc in tumorigenesis by inhibiting c-Myc-induced apoptosis via INK4a/ARF. *Genes Dev.* **13**, 2678–2690 (1999).
27. Eischen, C. M., Weber, J. D., Roussel, M. F., Sherr, C. J. & Cleveland, J. L. Disruption of the ARF-Mdm2-p53 tumour suppressor pathway in Myc-induced lymphomagenesis. *Genes Dev.* **13**, 2658–2669 (1999).
28. Verschuren, E. W., Klefstrom, J., Evan, G. I. & Jones, N. The oncogenic potential of Kaposi's sarcoma-associated herpesvirus cyclin is exposed by p53 loss *in vitro* and *in vivo*. *Cancer Cell* **2**, 229–241 (2002).
29. Tolbert, D., Lu, X., Yin, C., Tantama, M. & Van Dyke, T. p19(ARF) is dispensable for oncogenic stress-induced p53-mediated apoptosis and tumour suppression *in vivo*. *Mol. Cell Biol.* **22**, 370–377 (2002).
30. Khan, S. H., Moritsugu, J. & Wahl, G. M. Differential requirement for p19ARF in the p53-dependent arrest induced by DNA damage, microtubule disruption, and ribonucleotide depletion. *Proc. Natl Acad. Sci. USA* **97**, 3266–3271 (2000).
31. Rogoff, H. A. *et al.* Apoptosis associated with deregulated E2F activity is dependent on E2F1 and Atm/Nbs1/Chk2. *Mol. Cell Biol.* **24**, 2968–2977 (2004).
32. Liao, M. J., Yin, C., Barlow, C., Wynshaw-Boris, A. & Van Dyke, T. Atm is dispensable for p53 apoptosis and tumour suppression triggered by cell cycle dysfunction. *Mol. Cell Biol.* **19**, 3095–3102 (1999).
33. Conn, C. W., Lewellyn, A. L. & Maller, J. L. The DNA damage checkpoint in embryonic cell cycles is dependent on the DNA-to-cytoplasmic ratio. *Dev. Cell* **7**, 275–281 (2004).
34. Flores, E. R. *et al.* p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature* **416**, 560–564 (2002).
35. Urist, M. & Prives, C. p53 leans on its siblings. *Cancer Cell* **1**, 311–313 (2002).
36. Yang, A., Kaghad, M., Caput, D. & McKeon, F. On the shoulders of giants: p63, p73 and the rise of p53. *Trends Genet.* **18**, 90–95 (2002).
37. Senoo, M., Manis, J. P., Alt, F. W. & McKeon, F. p63 and p73 are not required for the development and p53-dependent apoptosis of T cells. *Cancer Cell* **6**, 85–89 (2004).
38. Irwin, M. S. *et al.* Chemoresistance linked to p73 function. *Cancer Cell* **3**, 403–410 (2003).
39. Bergamaschi, D. *et al.* p53 polymorphism influences response in cancer chemotherapy via modulation of p73-dependent apoptosis. *Cancer Cell* **3**, 387–402 (2003).
40. Gaidon, C., Lokshin, M., Ahn, J., Zhang, T. & Prives, C. A subset of tumour-derived mutant forms of p53 down-regulate p63 and p73 through a direct interaction with the p53 core domain. *Mol. Cell Biol.* **21**, 1874–1887 (2001).
41. Perez, D. & White, E. E1A sensitizes cells to tumour necrosis factor alpha by downregulating c-FLIP S. *J. Virol.* **77**, 2651–2662 (2003).
42. Klefstrom, J., Verschuren, E. W. & Evan, G. c-Myc augments the apoptotic activity of cytosolic death receptor signalling proteins by engaging the mitochondrial apoptotic pathway. *J. Biol. Chem.* **277**, 43224–43232 (2002).
43. Croxton, R., Ma, Y., Song, L., Haura, E. B. & Cress, W. D. Direct repression of the Mcl-1 promoter by E2F1. *Oncogene* **21**, 1359–1369 (2002).
44. Eischen, C. M. *et al.* Bcl-2 is an apoptotic target suppressed by both c-Myc and E2F-1. *Oncogene* **20**, 6983–6993 (2001).
45. Egle, A., Harris, A. W., Bouillet, P. & Cory, S. Bim is a suppressor of Myc-induced mouse B cell leukemia. *Proc. Natl Acad. Sci. USA* **101**, 6164–6169 (2004).
46. Hershko, T. & Ginsberg, D. Up-regulation of Bcl-2 homology 3 (BH3)-only proteins by E2F1 mediates apoptosis. *J. Biol. Chem.* **279**, 8627–8634 (2004).
47. Nahle, Z. *et al.* Direct coupling of the cell cycle and cell death machinery by E2F. *Nature Cell Biol.* **4**, 859–864 (2002).
48. Sears, R., Ohtani, K. & Nevins, J. R. Identification of positively and negatively acting elements regulating expression of the E2F2 gene in response to cell growth signals. *Mol. Cell Biol.* **17**, 5227–5235 (1997).
49. Matsumura, I., Tanaka, H. & Kanakura, Y. E2F1 and c-Myc in cell growth and death. *Cell Cycle* **2**, 333–338 (2003).

50. Leone, G. *et al.* Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol. Cell* **8**, 105–113 (2001).
51. Russell, J. L. *et al.* ARF differentially modulates apoptosis induced by E2F1 and Myc. *Mol. Cell. Biol.* **22**, 1360–1368 (2002).
52. Baudino, T. A. *et al.* Myc-mediated proliferation and lymphomagenesis, but not apoptosis, are compromised by E2f1 loss. *Mol. Cell* **11**, 905–914 (2003).
53. Conner, E. A. *et al.* Dual functions of E2F-1 in a transgenic mouse model of liver carcinogenesis. *Oncogene* **19**, 5054–5062 (2000).
54. Leone, G., DeGregori, J., Sears, R., Jakoi, L. & Nevins, J. R. Myc and Ras collaborate in inducing accumulation of active cyclin E/Cdk2 and E2F. *Nature* **387**, 422–426 (1997).
55. Dimri, G. P., Itahana, K., Acosta, M. & Campisi, J. Regulation of a senescence checkpoint response by the E2F1 transcription factor and p14(ARF) tumour suppressor. *Mol. Cell. Biol.* **20**, 273–285 (2000).
56. Damalas, A., Kahan, S., Shtutman, M., Ben-Ze'ev, A. & Oren, M. Deregulated beta-catenin induces a p53- and ARF-dependent growth arrest and cooperates with Ras in transformation. *EMBO J.* **20**, 4912–4922 (2001).
57. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
58. Campisi, J. Cellular senescence as a tumour-suppressor mechanism. *Trends Cell Biol.* **11**, S27–S31 (2001).
59. Shay, J. W. & Roninson, I. B. Hallmarks of senescence in carcinogenesis and cancer therapy. *Oncogene* **23**, 2919–2933 (2004).
60. Pearson, M. *et al.* PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. *Nature* **406**, 207–210 (2000).
61. Ferbeyre, G. *et al.* PML is induced by oncogenic ras and promotes premature senescence. *Genes Dev.* **14**, 2015–2027 (2000).
62. Itahana, K. *et al.* Control of the replicative life span of human fibroblasts by p16 and the polycomb protein Bmi-1. *Mol. Cell. Biol.* **23**, 389–401 (2003).
63. Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).
64. Paramio, J. M. *et al.* The ink4a/arf tumour suppressors cooperate with p21cip1/waf in the processes of mouse epidermal differentiation, senescence, and carcinogenesis. *J. Biol. Chem.* **276**, 44203–44211 (2001).
65. Elenbaas, B. *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15**, 50–65 (2001).
66. Horner, S. M., DeFilippis, R. A., Manuelidis, L. & DiMaio, D. Repression of the human papillomavirus E6 gene initiates p53-dependent, telomerase-independent senescence and apoptosis in HeLa cervical carcinoma cells. *J. Virol.* **78**, 4063–4073 (2004).
67. Wu, X. & Levine, A. J. p53 and E2F-1 cooperate to mediate apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 3602–3606 (1994).
68. Benanti, J. A. & Galloway, D. A. Normal human fibroblasts are resistant to RAS-induced senescence. *Mol. Cell. Biol.* **24**, 2842–2852 (2004).
69. Tuveson, D. A. *et al.* Endogenous oncogenic K-ras (G12D) stimulates proliferation and widespread neoplastic and developmental defects. *Cancer Cell* **5**, 375–387 (2004).
70. Guerra, C. *et al.* Tumour induction by an endogenous K-ras oncogene is highly dependent on cellular context. *Cancer Cell* **4**, 111–120 (2003).
71. Lin, A. W. & Lowe, S. W. Oncogenic ras activates the ARF-p53 pathway to suppress epithelial cell transformation. *Proc. Natl Acad. Sci. USA* **98**, 5025–5030 (2001).
72. Kelly-Spratt, K. S., Gurley, K. E., Yasui, Y. & Kemp, C. J. p19(ARF) suppresses growth, progression, and metastasis of Hras-driven carcinomas through p53-dependent and -independent pathways. *PLoS Biol.* **2**, E242 (2004).
73. Pelengaris, S., Littlewood, T., Khan, M., Elia, G. & Evan, G. Reversible activation of c-Myc in skin: induction of a complex neoplastic phenotype by a single oncogenic lesion. *Mol. Cell* **3**, 565–577 (1999).
74. Raff, M. C. Social controls on cell survival and cell death. *Nature* **356**, 397–400 (1992).
75. Askew, D., Ashmun, R., Simmons, B. & Cleveland, J. Constitutive c-myc expression in IL-3-dependent myeloid cell line suppresses cycle arrest and accelerates apoptosis. *Oncogene* **6**, 1915–1922 (1991).
76. Harrington, E. A., Bennett, M. R., Fanidi, A. & Evan, G. I. c-Myc-induced apoptosis in fibroblasts is inhibited by specific cytokines. *EMBO J.* **13**, 3286–3295 (1994).
77. Grossmann, J. Molecular mechanisms of 'detachment-induced apoptosis–Anoikis'. *Apoptosis* **7**, 247–260 (2002).
78. Vivanco, I. & Sawyers, C. L. The phosphatidylinositol 3-kinase AKT pathway in human cancer. *Nature Rev. Cancer* **2**, 489–501 (2002).
79. Plas, D. R., Rathmell, J. C. & Thompson, C. B. Homeostatic control of lymphocyte survival: potential origins and implications. *Nature Immunol.* **3**, 515–521 (2002).
80. Grad, J. M., Zeng, X. R. & Boise, L. H. Regulation of Bcl-xL: a little bit of this and a little bit of STAT. *Curr. Opin. Oncol.* **12**, 543–549 (2000).
81. LeRoith, D. & Helman, L. The new kid on the block(ade) of the IGF-1 receptor. *Cancer Cell* **5**, 201–202 (2004).
82. Christofori, G., Naik, P. & Hanahan, D. A second signal supplied by insulin-like growth factor II in oncogene-induced tumorigenesis. *Nature* **369**, 414–418 (1994).
83. Orsulic, S. *et al.* Induction of ovarian cancer by defined multiple genetic changes in a mouse model system. *Cancer Cell* **1**, 53–62 (2002).
84. Kauffmann-Zeh, A. *et al.* Suppression of c-Myc-induced apoptosis by Ras signalling through PI(3)K and PKB. *Nature* **385**, 544–548 (1997).
85. Wendel, H. G. *et al.* Survival signalling by Akt and eIF4E in oncogenesis and cancer therapy. *Nature* **428**, 332–337 (2004).
86. Zindy, F. *et al.* Arf tumour suppressor promoter monitors latent oncogenic signals *in vivo*. *Proc. Natl Acad. Sci. USA* **100**, 15930–15935 (2003).
87. Aslanian, A., Iaquinia, P. J., Verona, R. & Lees, J. A. Repression of the Arf tumour suppressor by E2F3 is required for normal cell cycle kinetics. *Genes Dev.* **18**, 1413–1422 (2004).
88. Jin, S. *et al.* CIAP1 and the serine protease HTRA2 are involved in a novel p53-dependent apoptosis pathway in mammals. *Genes Dev.* **17**, 359–367 (2003).
89. Zheng, T. S. *et al.* Deficiency in caspase-9 or caspase-3 induces compensatory caspase activation. *Nature Med.* **6**, 1241–1247 (2000).
90. Harlin, H., Reffey, S. B., Duckett, C. S., Lindsten, T. & Thompson, C. B. Characterization of XIAP-deficient mice. *Mol. Cell. Biol.* **21**, 3604–3608 (2001).
91. Sage, J., Miller, A. L., Perez-Mancera, P. A., Wysocki, J. M. & Jacks, T. Acute mutation of retinoblastoma gene function is sufficient for cell cycle re-entry. *Nature* **424**, 223–228 (2003).
92. Seoane, J., Le, H. V. & Massague, J. Myc suppression of the p21(Cip1) Cdk inhibitor influences the outcome of the p53 response to DNA damage. *Nature* **419**, 729–734 (2002).
93. Johnstone, R. W., Ruefli, A. A. & Lowe, S. W. Apoptosis: a link between cancer genetics and chemotherapy. *Cell* **108**, 153–164 (2002).
94. Chin, L. *et al.* Essential role for oncogenic Ras in tumour maintenance. *Nature* **400**, 468–472 (1999).
95. Felsner, D. W. & Bishop, J. M. Reversible tumorigenesis by MYC in haematopoietic lineages. *Mol. Cell* **4**, 199–207 (1999).
96. Gunther, E. J. *et al.* Impact of p53 loss on reversal and recurrence of conditional Wnt-induced tumorigenesis. *Genes Dev.* **17**, 488–501 (2003).
97. Moody, S. E. *et al.* Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. *Cancer Cell* **2**, 451–461 (2002).
98. Pakunlu, R. I., Cook, T. J. & Minko, T. Simultaneous modulation of multidrug resistance and antiapoptotic cellular defence by MDR1 and BCL-2 targeted antisense oligonucleotides enhances the anticancer efficacy of doxorubicin. *Pharm. Res.* **20**, 351–359 (2003).
99. Bykov, V. J. & Wiman, K. G. Novel cancer therapy by reactivation of the p53 apoptosis pathway. *Ann. Med.* **35**, 458–465 (2003).

Acknowledgements We thank members of the Lowe and Evan laboratories for discussions, and M. Hemann, D. Burgess and M. McCurrach for critical readings of the manuscript. S. L. is supported by an NRCR-AACR research professorship and grants from the National Institutes of Health and the Leukemia and Lymphoma Society, E. C. is supported by a postdoctoral training fellowship from the National Cancer Institute, and G. E. is supported by grants from the NIH, the Juvenile Diabetes Foundation and the Brain Tumor Society.

Competing interests statement The authors declare that they have no competing financial interests.

Cell-cycle checkpoints and cancer

Michael B. Kastan¹ & Jiri Bartek²

¹Department of Hematology-Oncology, St Jude Children's Research Hospital, 332 North Lauderdale Street, Memphis, Tennessee 38105, USA (e-mail: michael.kastan@stjude.org)

²Department of Cell Cycle and Cancer, Institute of Cancer Biology, Danish Cancer Society, Strandboulevarden 49, Copenhagen DK-2100, Denmark (e-mail: jb@cancer.dk)

All life on earth must cope with constant exposure to DNA-damaging agents such as the Sun's radiation. Highly conserved DNA-repair and cell-cycle checkpoint pathways allow cells to deal with both endogenous and exogenous sources of DNA damage. How much an individual is exposed to these agents and how their cells respond to DNA damage are critical determinants of whether that individual will develop cancer. These cellular responses are also important for determining toxicities and responses to current cancer therapies, most of which target the DNA.

The DNA contained in every mammalian cell is under constant attack by agents that can either directly damage one of its three billion bases or break the phosphodiester backbone on which the bases reside. For example, free oxygen radicals, which can cause both base damage and DNA breakage, arise as a consequence of normal cellular metabolism or can be created when the organism is exposed to external sources of ionizing radiation in the environment. Life on Earth has evolved to deal with both metabolic and external sources of DNA-damaging agents through the development of elegant mechanisms that repair damage to the DNA.

Cellular responses to DNA damage constitute one of the most important fields in cancer biology. First, damage to cellular DNA causes cancer. We know this from epidemiological studies¹, from animal models and from the observation that many human-cancer-susceptibility syndromes arise from mutations in genes involved in DNA-damage responses². Second, DNA damage is used to cure cancer. Most therapeutic modalities that we currently use to treat malignancies target the DNA, including radiation therapy and many chemo-therapeutic agents. Third, DNA damage is responsible for most of the side effects of therapy. Bone marrow suppression, gastrointestinal toxicities, and hair loss are all attributable to DNA-damage-induced cell death of proliferating progenitor cells in these tissues. So, from the perspective of cancer, DNA damage causes the disease, it is used to treat the disease, and it is responsible for the toxicity of therapies for the disease.

Among the mechanisms that cells have developed to cope with this constant attack on their DNA are elegant but not perfect DNA-repair processes. Because there are various types of DNA lesion that can occur, a variety of different repair mechanisms exist. In addition to directly repairing DNA breaks or adducts, cells respond to DNA damage by halting cell-cycle progression or by undergoing programmed cell death. Although we have a limited understanding of how the processes of cell-cycle arrest or apoptosis are coordinated with the process of DNA repair, such coordination must take place to optimize the outcome for the cell or the organism. In addition to damage to the DNA, cells must cope with other stresses, such as intermittent or prolonged exposure to abnormally low levels of oxygen or nutrients. Although cells use different aspects of the signalling pathways to deal with these types of change in their microenvironment, there are commonalities in the steps that cells use to deal with DNA damage.

The term 'cell-cycle checkpoint' refers to mechanisms by which the cell actively halts progression through the cell

cycle until it can ensure that an earlier process, such as DNA replication or mitosis, is complete³. Here, we focus on some of the mechanisms by which cells modulate progression through the cell cycle in the face of DNA damage and other stresses that affect DNA replication. Although we focus on signalling pathways that have been characterized in mammalian cells, lessons learned from studying lower eukaryotes (in particular, yeast), have been instructive and reflect the considerable evolutionary conservation of these pathways. Finally, current concepts about how sporadic or inherited mutations in genes in these pathways contribute to cancer development will be explored.

The signalling pathways

Signal initiation by different stresses

DNA can be damaged in a variety of ways. First, energy released by free oxygen radicals, generated either by normal metabolic processes or by exposure to an external source of ionizing radiation, can break the phosphodiester bonds in the backbone of the DNA helix. When two of these breaks are close to each other, but on opposite DNA strands, a double-strand break (DSB) is present in the DNA and the cell faces a particularly challenging situation for repair. Second, alkylating chemical moieties can modify purine bases and the size of the chemical adduct determines what repair process is used². Bifunctional alkylating chemicals can cause intra-strand or inter-strand crosslinks that require additional molecular interventions for them to be reversed. Third, inhibitors of DNA topoisomerases can lead to enhanced single or DSBs depending on which topoisomerase is inhibited and on the phase of the cell cycle⁴.

Each type of DNA damage requires a specific set of cellular responses to deal with the specific nature of the damage. Different mechanisms are required to repair the damage to the DNA backbone or to the DNA bases and the challenges of repairing the DNA can vary in the different phases of the cell cycle. To optimally repair DNA damage, the cell must also control other cellular processes before or during the repair, such as DNA replication or mitosis. Cells that are damaged when they are already in the middle of the process of DNA replication face particular challenges, but would still probably benefit from halting or slowing DNA replication until the damage has been repaired. So, there should be advantages for a eukaryotic cell to transiently halt progression through the cell cycle after DNA damage, which presumably include limiting heritable mutations in daughter cells and enhancing viability of the damaged cells.

Initiation of the activities of the PI(3)K (phosphatidylinositol-3-OH kinase)-like kinases (PIKKs), ATM (ataxia

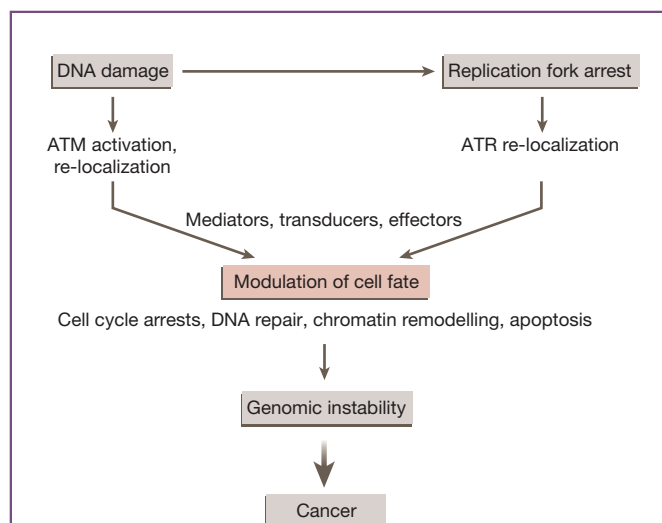


Figure 1 General scheme of responses to DNA damage or replication-fork arrest and the impact on cell fate, genomic instability and cancer development. Replication-fork arrest stimulates the initiation of cellular ATR activity, whereas DNA damage can directly activate ATM and can lead to replication-fork arrest, thereby also activating cellular ATR kinase. Once active, both the ATM and ATR kinases, functioning in combination with other proteins and substrates, help determine the outcome of the cell. If genomic instability ensues, this can contribute to cellular transformation.

telangiectasia mutated) and ATR (AMT- and Rad3-related) are the first steps characterized to date in the activation of signal transduction pathways that inhibit cell-cycle progression after DNA damage. The ATM kinase seems to primarily be activated following DNA damage whereas the ATR kinase seems to be critical for cellular responses to the arrest of DNA replication forks — the DNA structures formed during replication. This is the case whether the arrest of replication-fork progression is due to DNA damage or to other stresses^{5,6}. Because many types of DNA damage result both in the direct damage of the DNA and the arrest of DNA replication forks, ATM and ATR seem to participate together in many cellular-stress responses and complex joint responses must be coordinated (Fig. 1).

Signal initiation by ATM and ATR

To accomplish the physiological goal of minimizing the adverse effects of a stressful physiological situation, an arrest of cell-cycle progression should be engaged very rapidly after exposure to the stress. ATM and ATR are both extremely large (predicted molecular mass of 350 and 301 kilodaltons, respectively) protein kinases that phosphorylate numerous substrates to achieve their physiological goals⁷. It is a mechanistic challenge to tightly control the activities of these large kinases so that they do not stimulate growth-suppressive pathways in the absence of an appropriate stress, but can be activated instantaneously following exposure to the stress.

Patients, mice and cells lacking ATM are viable, suggesting that the ATM kinase is not essential for critical cellular functions such as normal cycle progression or cellular differentiation⁵. ATM kinase activity is minimal or low in unstressed cells and primarily is engaged to help cells deal with cellular stresses that affect DNA or chromatin structure. The identification of a single, major damage-induced phosphorylation site (serine 1981) led to the demonstration of a new mechanism of ATM regulation that permits a rapid and sensitive switch for checkpoint pathways⁸. In unstressed cells, ATM is present as a homodimer in which the kinase domain is physically blocked by its tight binding to an internal domain of the

protein surrounding serine 1981. The introduction of a DNA DSB leads to a conformational change in the ATM protein. This stimulates the kinase to phosphorylate serine 1981, causing the dissociation of the homodimer⁸. The activated ATM monomer can now phosphorylate its numerous substrates, whether they are nucleoplasmic, like p53, or at the sites of DNA breaks, like NBS1 (Nijmegen breakage syndrome 1), BRCA1 (breast cancer 1), and SMC1 (structural maintenance of chromosomes 1). The conformational change that induces the extremely rapid and extensive intermolecular autophosphorylation event in ATM does not seem to require the binding of the ATM dimer to sites of DNA damage, but instead results from some change in higher-order chromatin structure that the ATM dimer can sense at some distance away from the site of the DNA break⁸. The nature of this chromatin structure change and how ATM senses this change, including whether it is a direct or an indirect ‘sensing’ mechanism, remains to be discovered. Recent observations that the multiprotein complex MRE11(meiotic recombination 11)/RAD50/NBS1 (MRN) contributes to the activation of ATM after ionizing radiation — at least at low doses of ionizing radiation — may shed light on the mechanisms by which this activation process occurs^{9–13}.

The phosphorylation of substrates by the ATM kinase requires more than the dissociation of the ATM homodimer and the release of the blocked ATM kinase domain. The activated ATM monomer must also get to the sites in the cell where the substrates are present, such as at DNA breaks. It was recently demonstrated that MRE11 binds to ATM and enhances its ability to phosphorylate substrates *in vitro* in the presence of an appropriate mimic of DNA breaks¹⁴. Such a function is consistent with the observation that ATM can be activated by exposure to ionizing radiation in cells lacking NBS1 or BRCA1, but fails to migrate to the sites of DNA strand breaks. Once recruited to the DNA break, the activated ATM can then phosphorylate critical substrates like NBS1, BRCA1 and SMC1, which accumulate at these sites⁹. If ATM is activated but fails to get recruited to DNA breaks, as happens after ionizing radiation in cells lacking NBS1 or BRCA1, or if chromatin structure changes occur in the absence of DNA breaks, then ATM is still able to phosphorylate nucleoplasmic substrates, such as p53. Thus, DNA damage leads to ATM activation and substrate phosphorylation by two distinctive steps: (1) chromatin structure change induces intermolecular ATM autophosphorylation and homodimer dissociation; and (2) activated ATM monomer is recruited to its substrates, some of which localize to sites of DNA damage (Fig. 2). In this model, ATM activation and recruitment of MRN and BRCA1 to sites of DNA breaks are two distinct events.

Although the activity of ATM in *in vitro* kinase assays is increased after immunoprecipitation from irradiated cells, there is no measurable change in the kinase activity of ATR — even in the face of stresses where ATR is required to sustain normal cellular responses⁶. It seems that ATR kinase may be constitutively ready to phosphorylate substrates but have its cellular functions largely controlled by subcellular localization. ATR exists in a complex with the ATR-interacting protein (ATRIP), both before and after exposure to stresses such as ultraviolet irradiation^{15–17}. The observation that replication protein A (RPA), a single-stranded DNA (ssDNA)-binding protein involved in DNA replication, stimulates the *in vitro* binding of ATRIP to ssDNA led to a model in which ATR becomes localized to sites of replication-fork arrest by means of binding of ATRIP to RPA (ref. 16). In this model, any stimulus or stress that leads to an abnormal stretch of ssDNA, such as an arrested replication fork, would be decorated with RPA. The accumulation of RPA on the ssDNA would then lead to the recruitment of the ATRIP protein, and its heterodimeric partner, ATR. Once the active ATR kinase is localized to the ssDNA region, it can phosphorylate critical substrates, such as RAD17 and CHK1 (Fig. 2). Although the critical importance of RPA in the recruitment of ATR to ssDNA has been questioned¹⁷, the importance of this change in ATR localization is generally accepted.

As with ATM, the presence of an active ATR kinase in the cells is not sufficient for ATR to carry out its cellular functions. In addition to ATR, several other proteins and protein complexes must be recruited to the ssDNA site as well. These include the clamp-loading, RAD17-containing complex, RSR, which participates in the loading of the RAD9–RAD1–HUS1 (9–1–1) sliding clamp onto chromatin, and the clasp protein, which is independently recruited to chromatin^{18–20}. All these events are required for the phosphorylation of CHK1 by ATR and for the activation of the appropriate cell-cycle checkpoints (Fig. 2).

Whereas cells tolerate the absence of ATM, cells and animals lacking ATR seem to be non-viable^{21,22}. These observations suggest that ATR is probably required for normal progression through the cell cycle, even in the absence of cellular stress. Consistent with this concept, recent results suggest a critical role for ATR in the normal progression of DNA replication forks²³. Given the binding of ATR to regions of ssDNA, a role in normal replication-fork progression is perhaps not surprising.

In addition to its apparent roles in normal replication-fork progression, ATR is probably engaged in the cellular responses to many other types of cellular stress because so many of them affect the rate of replication-fork progression. ATR has been implicated in cellular responses to hypoxia²⁴ and to DNA-replication inhibitors^{16,25}. It is also critical for responses to DNA-damaging events that affect the progression of replication forks, particularly agents that introduce bulky DNA adducts, such as ultraviolet irradiation and alkylating agents or crosslinking agents. So, whereas ATM seems to become engaged in signalling pathways primarily following the introduction of DNA breaks, ATR has a critical role in virtually all cellular stress responses that share inhibition of replication-fork progression as a common mechanism. ATR even seems to be engaged in cellular responses to DNA breaks, possibly compensating for ATM: many ATM substrates eventually get phosphorylated after exposure to ionizing radiation in cells lacking ATM protein.

Transducing the signal

To efficiently spread the alert signal and orchestrate the global cellular response to DNA damage that is usually inflicted to only a few sites within the vast genome, the proximal checkpoint kinases ATM and ATR (ref. 26) cooperate closely with two other classes of proteins. These are the so-called checkpoint mediators (also known as adaptors)^{12,27,28} and the transducer kinases CHK1 and CHK2 (ref. 29; Fig. 3). Regulatory phosphorylations of the downstream checkpoint targets — diverse effector proteins that function at the interface between the cell-cycle, DNA-repair and cell-death machineries — may be carried out by the proximal kinases or transducer kinases alone. Alternatively, distinct residues of these same effectors are targeted by ATM/ATR and CHK1/CHK2, respectively^{26,29} (Fig. 3).

It is remarkable how rapidly (within seconds after the focal injury to DNA), the global checkpoint networks become activated and the local events at the damage site are coordinated with more distant cellular processes. For example, in response to only a few potentially harmful lesions, such as DSBs in proliferating cells, not only must the lesions be processed locally, but the whole cell must be alerted to delay the most vulnerable processes, such as DNA replication or initiation of chromosome segregation in a coordinated, 'pan-cellular' manner. Such speed and spatio-temporal coordination reflect the fact that the initial checkpoint responses operate through post-translational modifications, re-localizations, dynamic interactions, and changes of conformation and/or stability of pre-existing proteins, all phenomena that are jointly governed by these three classes of checkpoint regulators.

Exciting insights into these highly dynamic events have recently been obtained using new technologies for real-time imaging of fluorescently labelled checkpoint proteins in live cells, and phospho-specific antibodies that recognize proteins modified in response to DNA damage^{30–32}. Furthermore, the significance of proper checkpoint

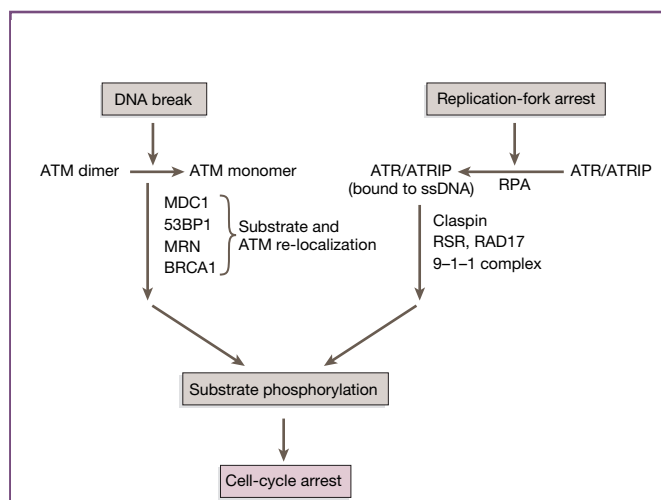


Figure 2 Scheme of mechanisms that lead to the induction of ATM- and ATR-directed cellular activities. DNA strand breaks lead to the dissociation of the inactive ATM dimer. The appropriate localization of both the ATM monomer and the ATM substrates is modulated by several proteins, including the MRN complex, MDC1, 53BP1, and Brca1. The ATR/ATRIP complex is recruited to sites of ssDNA, perhaps by RPA. Optimal substrate phosphorylation and the engagement of cell-cycle arrest depends on other proteins such as claspin, the RSR complex and the 9-1-1 complex. As illustrated in Fig. 1, these pathways may often operate in concert and there may be cross-talk between the pathway components shown here.

signalling for the prevention of cancer is underscored by the fact that most checkpoint kinases and mediators are either established or emerging tumour suppressors — gene products whose decreased expression or loss-of-function mutations contribute to tumorigenesis (Fig. 3). So, what are the roles and the underlying molecular mechanisms of action of the checkpoint mediators and the signal-transducing kinases?

The emerging role of checkpoint mediators

Although the precise mechanisms of action of this important class of checkpoint factors are largely unknown, they seem to modulate the activity of ATM/ATR, facilitate the interactions between ATM/ATR and their substrates, and in a broader sense 'mediate' spatio-temporal assembly of multiprotein complexes in the chromatin regions surrounding the sites of DNA damage. There are currently three known members of this class of checkpoint factors involved in the signalling by ATM; so far only one such protein is known to modulate the response by ATR (Fig. 3). As most of the mediators are initially recruited to sites of DNA damage and/or replication blockade independently of ATM and ATR, they might also be involved in 'sensing' such lesions. Alternatively, the mediator proteins could be recruited through their interaction with the candidate DNA-damage sensors^{33,34}.

The ATM-related mediators include MDC1 (mediator of DNA damage checkpoint 1; also known as NFB1), 53BP1 (p53 binding protein 1) and BRCA1 — large multi-domain proteins that contain two tandem BRCT (Brca1 carboxy-terminal) domains at their C terminus^{9,27,33,35–41}. Interestingly, the BRCT domains have recently been shown to serve as protein-phosphoprotein-binding modules^{42,43}, suggesting a possible mechanism for how the mediator proteins could promote the transient multiple interactions of checkpoint and repair proteins near the DNA-damage sites. Indeed, unlike the initial, largely ATM-independent, recruitment of the mediators to sites of DNA damage, their accumulation into the microscopically visible 'foci' depends on ATM-mediated phosphorylation of histone H2AX^{12,27,30–32}, a modification that marks chromatin regions spanning megadaltons of DNA flanking each DSB (ref. 44). The MDC1 protein

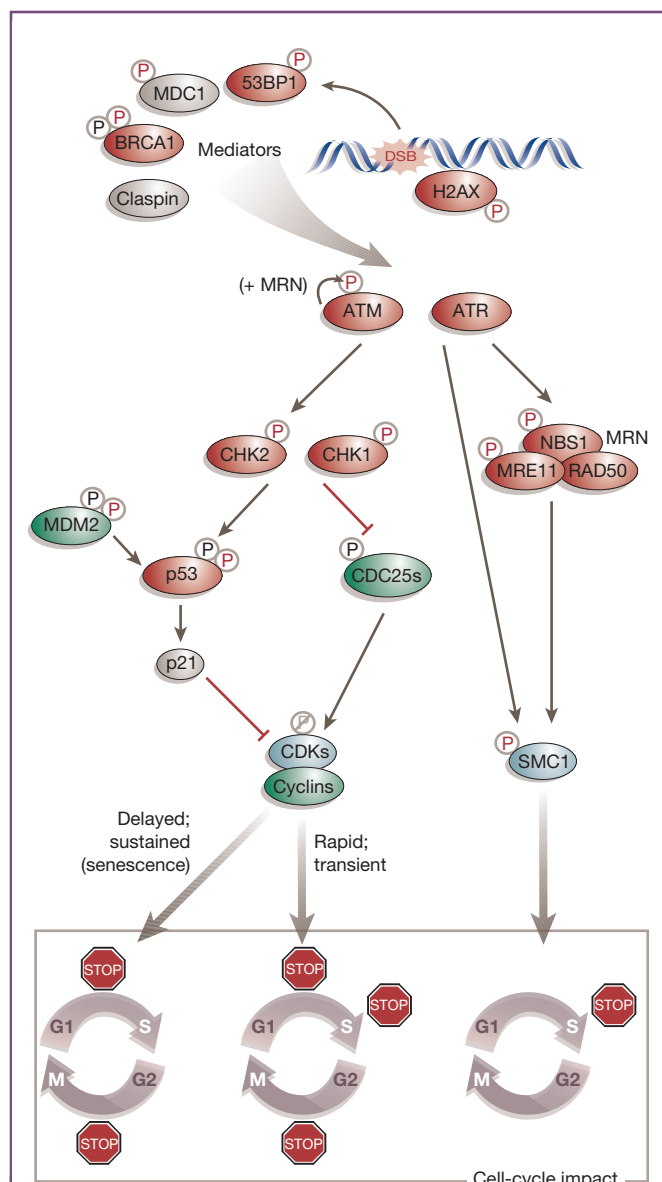


Figure 3 A simplified scheme of cell-cycle checkpoint pathways induced in response to DNA damage (here DSBs), with highlighted tumour suppressors shown in red and proto-oncogenes shown in green. The proximal checkpoint kinases ATM and ATR phosphorylate diverse components of the network, either directly (red 'P') or through the transducing kinases CHK2 and CHK1 (black 'P'). (For simplicity, some candidate damage sensors and several ATM/ATR and CHK1/CHK2 substrates have been omitted.) The BRCA1 protein also contributes to cell-cycle arrest and DNA repair by homologous recombination, whereas p53 controls genes involved in cell death and DNA-repair mechanisms. The cell-cycle phase and the duration of the blockade affected by the effector pathways are indicated, including the potential permanent arrest (senescence), as mediated by p53. The global checkpoint network regulated by ATM/ATR and CHK2/CHK1 also affects cellular responses other than cell cycle progression, including DNA repair, transcription, chromatin assembly and cell death.

functions as a molecular bridge between the phosphorylated H2AX (γ -H2AX) and the NBS1 component of the MRN complex³¹, and helps provide a platform for a myriad of dynamic interactions for these and additional checkpoint and DNA-repair proteins (including the activated ATM and BRCA1) within the vicinity of the damage sites. Although the mediator proteins are unlikely to initially target the activated ATM to sites of DNA damage (this might be the role of the candidate damage sensors such as the MRN complex)^{9–11,13,45,46}, the sustained multiprotein interactions mediated by MDC1, 53BP1

and BRCA1 seem to facilitate ATM signalling and the processing/repair of the lesions, thereby contributing to the biological outcome of the checkpoint responses (Fig. 2)^{9,12,27,33,35–41}. Consistent with this concept, mammalian cells that lack any of these three mediators show enhanced sensitivity to DNA-damaging agents such as ionizing radiation, and impaired intra-S-phase and G2/M cell-cycle checkpoints.

Reminiscent of the roles of MDC1, 53BP1 and BRCA1 in proper localization, timing and velocity of the ATM-controlled signalling, the ATR-controlled checkpoint signalling, at least towards the Chk1 kinase that is activated by ATR, relies on claspin²⁸. Claspin is a mediator/adaptor protein that is structurally unrelated to the mediators involved in response to DSBs. Claspin selectively interacts with chromatin structures created by active replication forks, and is required for ATR-mediated phosphorylation, and so for proper activation, of CHK1 (Fig. 2)^{28,47}.

Effector kinases CHK1 and CHK2

Prominent among the substrates of the apical checkpoint kinases ATM and ATR are the checkpoint-transducer serine/threonine kinases (also known as effector kinases) CHK2 and CHK1 (ref. 29). Despite some 'cross-talk' between ATM and CHK1, the ATM- and ATR-mediated phosphorylations trigger preferentially the activation of CHK2 and CHK1 (Fig. 3), respectively²⁹. Given that the ATM–CHK2 and ATR–CHK1 signalling modules share many substrates among the checkpoint effector proteins^{26,29}, it is striking that ATM and CHK2 are dispensable for pre-natal development, whereas complete absence of either ATR or CHK1 results in early embryonic lethality^{26,29}. As mentioned above for ATR, a plausible explanation for such a fundamental biological difference emerges from recent evidence that supports a role for ATR–CHK1 in the regulation of some essential processes during unperturbed cell cycles, including the control of DNA replication^{23,48} or initiation of mitotic events on centrosomes⁴⁹. More mechanistic insights into how the checkpoint kinases ATM and ATR — in concert with the extremely mobile messenger kinases CHK2/CHK1 (refs 29, 30) — trigger cell-cycle delays at various transitions of the cell-division cycle is the subject of the next section.

Affecting the cell cycle

During unperturbed proliferation, mammalian cells can only withdraw from the cell cycle on experiencing growth-factor deprivation or growth inhibitory signals in early-to-mid G1 phase (see review in this issue by Massagué, page 298). This is before the cells pass through the RB (retinoblastoma protein)/E2F (transcription factor)-controlled restriction point, after which they are committed to a round of DNA replication and cell division^{50,51}. However, the ATM/ATR–CHK2/CHK1-controlled checkpoint network response to genotoxic stress can transiently delay cell-cycle progression in G1, S or G2 phases, or even impose prolonged, durable cell-cycle arrests in either G1 or G2, before entry into the subsequent S phase or mitosis, respectively. Given the critical significance of error-free DNA replication and chromosome segregation for the maintenance of genomic integrity and the prevention of cancer, it is not surprising that these most vulnerable stages of the cell-division cycle are protected by a wider spectrum of checkpoint effector mechanisms, the identity of which are briefly discussed below.

The G1 and G1/S checkpoint responses

The dominant checkpoint response to DNA damage in mammalian cells traversing through G1 is the ATM(ATR)/CHK2(CHK1)–p53/MDM2–p21 pathway (Fig. 3), which is capable of inducing sustained, and sometimes even permanent G1 arrest^{7,29,52}. Although the expression of ATM and CHK2 is relatively constant during the cell cycle, the concentrations of ATR and CHK1 are low in the early-to-mid G1, and their activities become important only closer to the G1/S transition. ATM/ATR directly phosphorylate the p53 transcription factor within its amino-terminal transactivation domain, particularly

on serine 15. Threonine 18 and serine 20 in the same domain, along with probably some additional p53 sequence(s), are also targeted by CHK1/CHK2 (refs 7, 26, 29, 52, 53). In addition, the ubiquitin ligase MDM2 that normally binds p53 and ensures rapid p53 turnover, is targeted after DNA damage by ATM/ATR (ref. 54), as well as by CHK2/CHK1 (N. Motoyama, personal communication). These modifications of p53 and Mdm2 contribute to the stabilization and accumulation of the p53 protein, as well as to its increased activity as a transcription factor. The key transcriptional target of p53 is the p21CIP1/WAF1 inhibitor of cyclin-dependent kinases^{7,52}, which silences the G1/S-promoting cyclin E/Cdk2 kinase and thereby causes a G1 arrest. This leads not only to the inability to initiate DNA synthesis, but it also preserves the RB/E2F pathway in its active, growth-suppressing mode, thereby causing a sustained G1 blockade (see also review in this issue by Massagué, page 298). Thus, the G1 checkpoint response targets two critical tumour suppressor pathways, governed by p53 and pRB. These are arguably the two mechanisms that are most commonly deregulated in human cancer^{7,50–52}.

In late G1, as part of the activated E2F-dependent S-phase-promoting transcriptional programme, the expression of ATR and CHK1 increases. Cyclins E and A, and the activator of the cyclin E(A)/CDK2 kinase — the CDC25A phosphatase — are also induced in late G1. The ATR/CHK1 module (but not ATM/CHK2), through moderate constitutive phosphorylation of CDC25A on its several serine residues, then maintains an appropriate abundance of CDC25A through its ubiquitin-dependent, proteasome-mediated turnover during unperturbed proliferation^{29,55}. In response to genotoxic stress, this physiologically operating mechanism becomes enhanced through increased activity of CHK1 and CHK2, leading to downregulation of CDC25A and consequently to the inhibition of cyclin E(A)/CDK2 complexes^{29,55,56}. Importantly, despite the simultaneous phosphorylation of CDC25A and p53 by checkpoint kinases (Fig. 3), the impact of these events on cell-cycle machinery is faster in the CDC25A-degradation cascade that unlike the slower-operating p53 pathway, does not require the transcription and accumulation of newly synthesized proteins. Thus, the CHK1/CHK2–CDC25A checkpoint is implemented rapidly, independently of p53, and it delays the G1/S transition only for a few hours, unless the sustained p53-dependent mechanism prolongs the G1 arrest.

The S-phase checkpoint pathways

The intra-S-phase checkpoint network activated by genotoxic insults causes largely transient, reversible inhibition of the firing from those origins of DNA replication that have not yet been initiated. It seems that there are at least two parallel branches of this checkpoint that slow down the ongoing DNA synthesis, both of which are controlled by the ATM/ATR signalling machinery. One of these effector mechanisms operates through the CDC25A-degradation cascade described in the previous section. The inhibition of CDK2 activity downstream of this pathway blocks the loading of CDC45 onto chromatin. CDC45 is a protein required for the recruitment of DNA polymerase α into assembled pre-replication complexes, so the inhibition of CDK2 activity prevents the initiation of new origin firing^{29,55}.

The other branch of the intra-S-phase checkpoint reflects the impact of ATM-mediated phosphorylations of NBS1 on several sites, in particular serine 343 (refs 7, 26) and serines 957 and 966 of the cohesin protein SMC1 (refs 9, 57, 58). A better mechanistic understanding of this pathway, whose proper function also depends on BRCA1 and FANCD2 (Fanconi anaemia, complementation group D2) proteins^{9,59,60}, should be particularly rewarding because the observed hypersensitivity to radiation in cells that are defective in NBS1 or SMC1 seems attributable to the inability of ATM to phosphorylate the two critical residues of the SMC1 effector⁹.

The concept of the two above-mentioned parallel effector branches of the intra-S-phase checkpoint has been documented for responses to both ionizing radiation (ref. 61) and to ultraviolet light⁶². Whether the recently reported targeting of CDC7 — another

kinase involved in regulation of DNA replication through an ATR-dependent DNA-damage response⁶³ — represents yet another parallel mechanism to delay DNA synthesis remains to be established.

Apart from the inhibition of replication-origin firing, another critical function provided by S-phase checkpoints (particularly the so-called 'replication checkpoint' activated by stalled replication) is to protect the integrity of the stalled replication forks. Such maintenance of fork stability, achieved through yet-to-be discovered effector mechanisms, helps prevent the conversion of primary lesions into DNA breaks and facilitates the subsequent recovery of DNA replication^{18,55}.

The G2 checkpoint

The G2 checkpoint (also known as the G2/M checkpoint) prevents cells from initiating mitosis when they experience DNA damage during G2, or when they progress into G2 with some unrepaired damage inflicted during previous S or G1 phases^{64,65}. The accumulation of cells in G2 may also reflect a contribution of the so-called DNA-replication checkpoint (often referred to as the S/M checkpoint) that may sense some of the persistent DNA lesions from the previous S phase as being inappropriately or not fully replicated DNA.

The critical target of the G2 checkpoint is the mitosis-promoting activity of the cyclin B/CDK1 kinase, whose activation after various stresses is inhibited by ATM/ATR, CHK1/CHK2 and/or p38-kinase-mediated subcellular sequestration, degradation and/or inhibition of the CDC25 family of phosphatases that normally activate CDK1 at the G2/M boundary^{56,65–68}. In addition, other upstream regulators of CDC25C and/or cyclin B/CDK1, such as the Polo-like kinases PLK3 and PLK1 seem to be targeted by DNA-damage-induced mechanisms⁶⁵. Analogous to the role of the checkpoint mediators in the S-phase checkpoint, 53BP1 and BRCA1 are also involved in the regulation of the G2-checkpoint responses^{39,41,69}.

The maintenance phase of the G2 checkpoint probably partly relies on the transcriptional programmes regulated by BRCA1 and p53, leading to the upregulation of cell-cycle inhibitors such as the CDK inhibitor p21, GADD45a (growth arrest and DNA-damage-inducible 45 α) and 14–3–3 sigma proteins^{65,70}. The fact that even tumours defective in other checkpoints, such as those with mutant p53, tend to selectively accumulate in G2 after DNA damage, indicates that p53-independent mechanisms are sufficient to sustain the G2 arrest. At the same time, this phenomenon has inspired efforts to interfere with the G2 checkpoint as a potential strategy to sensitize cancer cells, which are deficient in their G1/S checkpoint pathways, to radiation- or drug-induced DNA damage⁷¹.

Impacting cancer

As the checkpoint and repair pathways facilitate cellular responses to DNA damage, and because there is significant data suggesting that DNA damage from both endogenous and exogenous sources is a major contributor to the development of human cancers, it is reasonable to speculate that alterations in these pathways increase the risk of cancer developing. Data from both animal models and humans strongly support this concept (Table 1). Cells with an intact DNA-damage response frequently arrest or die in response to DNA damage, thus reducing the likelihood of progression to malignancy. Mutations in apoptosis, DNA-damage responses or in mitotic-checkpoint pathways, however, can permit the survival or the continued growth of cells with genomic abnormalities, thereby enhancing the chance of malignant transformation. Although many of the DNA-damage response factors described above have been classified as tumour suppressor genes and oncogenes (see also Fig. 3), the dysfunction of these pathways has not been linked to cancer development in all cases. Sorting out which pathway steps are important in affecting the predisposition to malignancies versus those that are not may provide invaluable insights into the mechanisms responsible for human tumorigenesis. Although germline mutations in mice and humans are used to identify the genes and pathway steps that predispose

Table 1 Human cancer susceptibility linked to DNA-damage response

Disease	Gene	Number of mutant alleles inherited	Cancer predisposition	Comments
Ataxia-telangiectasia (A-T)	<i>ATM</i>	2	Leukaemia, lymphoma	Most mutations result in null protein phenotype
Nijmegen breakage syndrome (NBS)	<i>NBS1</i>	2	Leukaemia, lymphoma	Fragment of NBS1 protein still expressed in some cell types
A-T-like disorder (ATLD)	<i>Mre11</i>	2	Leukaemia, lymphoma	Hypomorphic mutations in <i>Mre11</i>
Fanconi's anaemia (FA)	<i>FancD2, Brca2</i> (also known as <i>FancD1</i>)	2	Acute myelogenous leukaemias	Other FA genes not directly implicated in checkpoints; <i>Brca2</i> — hypomorphic
Familial breast, ovarian carcinoma syndrome	<i>Brca1, Brca2</i>	1	Breast, ovarian, scattered others	
Li-Fraumeni syndrome	<i>p53, CHEK2</i>	1	Sarcomas, leukaemias, brain tumours, adrenal tumours, others	

This list does not include syndromes resulting from DNA-repair defects, which includes xeroderma pigmentosum, hereditary non-polyposis colon cancers, Bloom's syndrome and other Fanconi's anaemia complementation groups.

animals to acquiring tumours, it is likely that dysfunction of these steps is also critically important in the development of sporadic tumours, which constitute most human cancers.

DNA-damage signal transducers and cancer

Loss of ATM strongly predisposes both humans and mice to lymphoma development⁵, and to a lesser degree to other malignancies²⁶. Because the deletion of the Rad52 protein, which is required for homologous recombination, significantly reduces lymphoma development in ATM-deficient mice, it has been suggested that excessive recombination is an important contributor to tumorigenesis in ataxia telangiectasia⁷². Similarly, patients with mutations in NBS1 or MRE11 are predisposed to develop cancer^{73–75}.

In contrast to the disruption of both alleles of ATR causing embryonic lethality in mice²², a human disease, Seckel syndrome, was recently associated with hypomorphic mutations in ATR that lead to low levels of ATR expression⁷⁶. Interestingly, although these patients show growth retardation, dwarfism, microcephaly and mental retardation, and their cells show chromosome instability after mitomycin C exposure⁷⁷, a high incidence of malignancies is not thought to be a prominent part of this inherited syndrome. However, ATR haploinsufficiency enhances tumorigenesis in mice that are defective for DNA-mismatch repair⁷⁸.

Certain mutations in additional components of these signalling pathways also lead to cancer predisposition. Mice lacking either *H2AX* (refs 79, 80) or *53BP1* (ref. 40) show cell-cycle checkpoint defects and cancer predisposition. Even haploinsufficiency for *H2AX* results in detectable genomic instability and enhanced tumour susceptibility in the absence of p53 (refs 79, 80). Although *H2AX* maps to a cytogenetic region commonly altered in human cancers, 11q23, it is not clear whether *H2AX* abnormalities contribute to human cancer. Although Mdc1 seems to be required for cell-cycle checkpoint function^{27,81}, mutations in the gene have not yet been linked to enhanced tumour development in mice or humans.

The homozygous-deficient state cannot be tested, but *Chk1* heterozygosity modestly enhances the tumorigenic phenotype of *Wnt1* transgenic mice⁸². As the tumours in these mice did not lose the other allele of *Chk1*, a haploinsufficient tumour suppressor mechanism was suggested. Potential mechanisms underlying the haploinsufficient phenotype were studied using generations of mice in which *Chk1* was conditionally disrupted in mammary epithelial cells. These cells showed inappropriate S-phase entry, accumulation of DNA damage during replication and inappropriate mitotic entry⁸³. These observations suggest that checkpoint defects associated with *Chk1* haploinsufficiency can contribute to tumorigenesis. *Chk2*^{−/−} mice do not spontaneously develop tumours⁸⁴, but a lack of Chk2 enhances skin tumorigenesis induced by carcinogen exposure. As inherited mutations in one allele of *CHEK2* can be found in some families with the extremely cancer-prone Li-Fraumeni syndrome⁸⁵, and *CHEK2* variants predispose individuals to breast and prostate cancer²⁹, *CHEK2* seems to be a complex tumour suppressor gene.

From BRCA to p53 to cancer

The inheritance of a single mutated allele of either *BRCA1* or *BRCA2* markedly increases the incidence of breast and ovarian cancers in women⁸⁶. As the tumours from these individuals virtually always lose the second allele, both *BRCA* genes conform to the classic pattern of tumour suppressor genes⁸⁷. It is now clear that both *BRCA* gene products participate in cellular responses to DNA damage, but they seem to have distinct roles. As described above, BRCA1 is a target of the ATM, ATR and CHK2 kinases and is required for cell-cycle checkpoint responses in S phase and G2/M (ref. 69). BRCA1 also localizes to sites of DNA breakage, interacts with chromatin remodelling proteins and has been implicated in transcriptional control⁸⁷. Which of these, or other suggested functions of BRCA1, are critical for tumour suppression and which explain the relative specificity for breast and ovarian cancers associated with its mutation remain to be clarified. Mouse models have suggested complex answers to this question. Because bi-allelic disruption of *Brca1* in the mouse results in embryonic lethality, tissue targeting and conditional disruptions have been used to assess the function of Brca1 (ref. 88). Increased

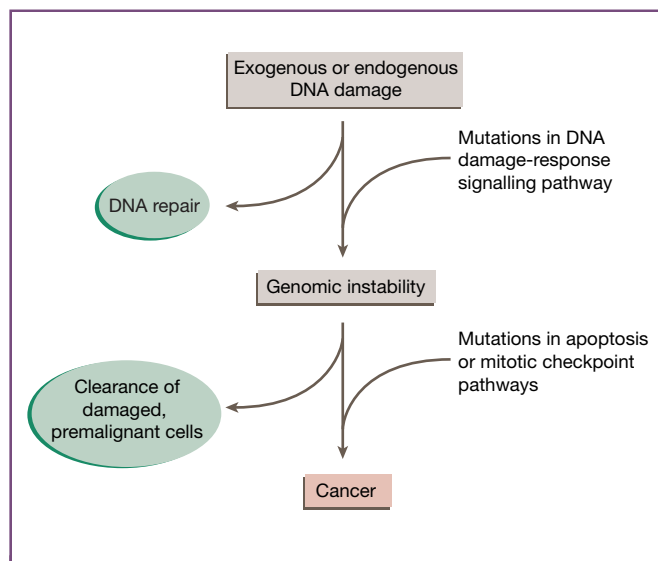


Figure 4 Schematic representation of two main steps that contribute to a spectrum of mutations leading to cancer development. If DNA damage is repaired efficiently, the likelihood of tumour development is low. If cells have mutations in DNA-damage-response signalling pathways — either sporadic or inherited — this will lead to enhanced genomic abnormalities. Cells with damaged DNA frequently arrest or do not survive, thus reducing the probability that they will progress to malignancy. Mutations in apoptosis pathways, DNA-damage, DNA-repair or mitotic-checkpoint pathways can permit the survival or continued growth of cells with genomic abnormalities, thus enhancing the likelihood of malignant transformation.

mammary and lymphoma carcinogenesis is seen in combination with p53 disruption. This suggests that p53-mediated apoptosis normally eliminates cells with enhanced DNA damage associated with *Brca1* disruption⁸⁹. Disruption of *Chk2* is less potent at enhancing the *Brca1* effects than disruption of p53, suggesting that some of the p53 tumour suppressor functions are retained in the absence of *Chk2* (ref. 90).

BRCA2 binds directly to the RAD51 recombinase and has been linked to the S-phase checkpoint and to homologous recombination functions⁹¹. A direct link between BRCA2 and the cancer-prone Fanconi's anaemia syndrome arose when patients with the Fanconi's D1 complementation group turned out to harbour biallelic hypomorphic mutations in the *BRCA2* gene⁹². In addition, BRCA1 and the Fanconi's D2 protein interact in DNA-damage signalling pathways (see section 'The S-phase checkpoint pathways' above). Although mice bearing mutations in the Fanconi's A or C genes show chromosome instability and defective germ-cell development, they do not spontaneously develop cancer⁹³. In contrast, mice lacking the Fanconi's D2 gene and *Brca2* hypomorphic mice develop epithelial cancers, such as breast, ovarian and liver cancer. Although mice with heterozygous mutations in *Brca2* do not develop tumours at an increased frequency, mice with homozygous truncations of *Brca2* develop thymic lymphomas. Growth arrest and unstable chromosome structure induced by *Brca2* truncation are relieved when cell-cycle checkpoints that are responsive to mitotic spindle disruption are inactivated⁹⁴. This suggests that inactivating mutations in mitotic checkpoint genes might cooperate with *Brca2* deficiency in the pathogenesis of inherited breast cancer and potentially other diseases of chromosomal instability, such as Blooms syndrome or Fanconi anaemia. This concept of mutations that cooperate with checkpoint or repair defects to enhance tumour development is likely to be a recurring theme in future studies (Fig. 4).

The BRCA stories suggest that genetic instability caused by altered DNA-damage response pathways may not be sufficient to lead to cancer development, and that cooperating mutations must be present to facilitate continued growth or viability of pre-malignant cells. Similarly, mice bearing hypomorphic mutations in the *Mre11* genes show pronounced chromosomal instability but are not prone to malignancy⁹⁵. However, tumour formation in these mice on a p53-heterozygote background is significantly enhanced, suggesting that the combination of genomic instability and cell-cycle checkpoint defects is a significant risk factor for tumour development. One recent report demonstrated that mice bearing a mutation in p53 that was defective in apoptosis, but retained some cell-cycle checkpoint function, was markedly less prone to tumour development than p53-null mice. This surprising result suggests that the growth arrest and chromosome stability functions of p53 provide tumour suppressor function even in the absence of its role in apoptosis⁹⁶. Finally, the recent observation that ATR haploinsufficiency increases tumorigenesis on a background of mismatch repair deficiency⁷⁸ may presage a flurry of new insights into how heterozygous mutations, although seemingly innocuous on their own, can enhance tumour formation when present in certain combinations, such as those controlling checkpoint responses and repair abilities.

Future directions

Animal models and human-cancer-susceptibility syndromes will continue to teach us about the physiological roles of the genes and pathways involved in DNA-damage responses. Many questions remain, such as how the cross-talk between the signalling pathways discussed here, and the processes of DNA repair and apoptosis operate. As these pathways seem to be major determinants of cellular responses to the types of cytotoxic agent that we use to treat tumours, these insights may teach us new ways to more effectively treat tumours. Similarly, because these response pathways seem to be major protectors from cancer development, the study of these pathways could lead to effective and new approaches to the reduction

of cancer development. In addition to the prevention of cancer and more effective treatment of malignancies, insights into the mechanisms involved in these response pathways may even shed light on the processes of aging and senescence. □

doi:10.1038/nature03097

- Doll, R. & Peto, R. The causes of cancer in the United States today. *J. Natl Cancer Inst.* **66**, 1192–1308 (1981).
- Ford, J. M. *Clinical Oncology* (eds Abeloff, M. et al.) Ch. 11, 191–205 (Elsevier, Philadelphia, 2004).
- Hartwell, L. H. & Weinert, T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science* **246**, 629–634 (1989).
- Froelich-Ammon, S. J. & Oshero, N. Topoisomerase poisons: harnessing the dark side of enzyme mechanism. *J. Biol. Chem.* **270**, 21429–21432 (1995).
- Shiloh, Y. & Kastan, M. B. ATM: genome stability, neuronal development, and cancer cross paths. *Adv. Cancer Res.* **83**, 209–254 (2001).
- Abraham, R. T. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes Dev.* **15**, 2177–2196 (2001).
- Kastan, M. B. & Lim, D.-S. The many substrates and functions of ATM. *Mol. Cell Biol.* **1**, 179–186 (2000).
- Bakkenist, C. J. & Kastan, M. B. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* **421**, 499–506 (2003).
- Kitagawa, R., Bakkenist, C. J., McKinnon, P. J. & Kastan, M. B. Phosphorylation of SMC1 is a critical downstream event in the ATM–NBS1–BRCA1 pathway. *Genes Dev.* **18**, 1423–1438 (2004).
- Uziel, T. et al. Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J.* **22**, 5612–5621 (2003).
- Carson, C. T. et al. The Mre11 complex is required for ATM activation and the G2/M checkpoint. *EMBO J.* **22**, 6610–6620 (2003).
- Mochan, T. A., Venere, M., DiTullio, R. A. Jr & Halazonetis, T. D. 53BP1, an activator of ATM in response to DNA damage. *DNA Repair (Amst)* **3**, 945–952 (2004).
- Horejsi, Z. et al. Distinct functional domains of Nbs1 modulate the timing and magnitude of ATM activation after low doses of ionizing radiation. *Oncogene* **23**, 3122–3127 (2004).
- Lee, J. H. & Paull, T. T. Direct activation of the ATM protein kinase by the Mre11/Rad50/Nbs1 complex. *Science* **304**, 93–96 (2004).
- Cortez, D., Guntuku, S., Qin, J. & Elledge, S. J. ATR and ATRIP: partners in checkpoint signaling. *Science* **294**, 1713–1716 (2001).
- Zou, L. & Elledge, S. J. Sensing DNA damage through ATRIP recognition of RPA–ssDNA complexes. *Science* **300**, 1542–1548 (2003).
- Unsal-Kacmaz, K. & Sancar, A. Quaternary structure of ATR and effects of ATRIP and replication protein A on its DNA binding and kinase activities. *Mol. Cell Biol.* **24**, 1292–1300 (2004).
- Osborn, A. J., Elledge, S. J. & Zou, L. Checking on the fork: the DNA-replication stress-response pathway. *Trends Cell Biol.* **12**, 509–516 (2002).
- Ellison, V. & Stillman, B. Opening of the clamp: an intimate view of an ATP-driven biological machine. *Cell* **106**, 655–660 (2001).
- Lin, S. Y., Li, K., Stewart, G. S. & Elledge, S. J. Human Claspin works with BRCA1 to both positively and negatively regulate cell proliferation. *Proc. Natl Acad. Sci. USA* **101**, 6484–6489 (2004).
- Zou, L., Cortez, D. & Elledge, S. J. Regulation of ATR substrate selection by Rad17-dependent loading of Rad9 complexes onto chromatin. *Genes Dev.* **16**, 198–208 (2002).
- Brown, E. J. & Baltimore, D. Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. *Genes Dev.* **17**, 615–628 (2003).
- Shechter, D., Costanzo, V. & Gautier, J. ATR and ATM regulate the timing of DNA replication origin firing. *Nature Cell Biol.* **6**, 648–655 (2004).
- Hammond, E. M., Denko, N. C., Dorie, M. J., Abraham, R. T. & Giaccia, A. J. Hypoxia links ATR and p53 through replication arrest. *Mol. Cell Biol.* **22**, 1834–1843 (2002).
- Hekmat-Nejad, M., You, Z., Yee, M. C., Newport, J. W. & Cimprich, K. A. Xenopus ATR is a replication-dependent chromatin-binding protein required for the DNA replication checkpoint. *Curr. Biol.* **10**, 1565–1573 (2000).
- Shiloh, Y. ATM and related protein kinases: safeguarding genome integrity. *Nature Rev. Cancer* **3**, 155–168 (2003).
- Stucki, M. & Jackson, S. P. MDC1/NFBD1: a key regulator of the DNA damage response in higher eukaryotes. *DNA Repair (Amst)* **3**, 953–957 (2004).
- Chini, C. C. & Chen, J. Claspin, a regulator of Chk1 in DNA replication stress pathway. *DNA Repair (Amst)* **3**, 1033–1037 (2004).
- Bartek, J. & Lukas, J. Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell* **3**, 421–429 (2003).
- Lukas, C., Falck, J., Bartkova, J., Bartek, J. & Lukas, J. Distinct spatiotemporal dynamics of mammalian checkpoint regulators induced by DNA damage. *Nature Cell Biol.* **5**, 255–260 (2003).
- Lukas, C. et al. Mdc1 couples DNA double-strand break recognition by Nbs1 with its H2AX-dependent chromatin retention. *EMBO J.* **23**, 2674–2683 (2004).
- Celeste, A. et al. Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5**, 675–679 (2003).
- Goldberg, M. et al. MDC1 is required for the intra-S-phase DNA damage checkpoint. *Nature* **421**, 952–956 (2003).
- Huyen, Y. et al. Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks. *Nature* doi: 10.1038/nature03114 (in the press).
- Lou, Z., Minter-Dykhouse, K., Wu, X. & Chen, J. MDC1 is coupled to activated CHK2 in mammalian DNA damage response pathways. *Nature* **421**, 957–961 (2003).
- Shang, Y. L., Boder, A. J. & Chen, P. L. NFBD1, a novel nuclear protein with signature motifs of FHA and BRC1, and an internal 41-amino acid repeat sequence, is an early participant in DNA damage response. *J. Biol. Chem.* **278**, 6323–6329 (2003).
- Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. & Elledge, S. J. MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature* **421**, 961–966 (2003).
- Xu, X. & Stern, D. F. NFBD1/KIAA0170 is a chromatin-associated protein involved in DNA damage signaling pathways. *J. Biol. Chem.* **278**, 8795–8803 (2003).

39. DiTullio, R.A. Jr *et al.* 53BP1 functions in an ATM-dependent checkpoint pathway that is constitutively activated in human cancer. *Nature Cell Biol.* **4**, 998–1002 (2002).
40. Ward, I. M., Minn, K., van Deursen, J. & Chen, J. p53 Binding protein 53BP1 is required for DNA damage responses and tumor suppression in mice. *Mol. Cell Biol.* **23**, 2556–2563 (2003).
41. Wang, B., Matsuoka, S., Carpenter, P. B. & Elledge, S. J. 53BP1, a mediator of the DNA damage checkpoint. *Science* **298**, 1435–1438 (2002).
42. Manke, I. A., Lowery, D. M., Nguyen, A. & Yaffe, M. B. BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science* **302**, 636–639 (2003).
43. Yu, X., Chini, C. C., He, M., Mer, G. & Chen, J. The BRCT domain is a phospho-protein binding domain. *Science* **302**, 639–642 (2003).
44. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks *in vivo*. *J. Cell Biol.* **146**, 905–916 (1999).
45. D'Amours, D. & Jackson, S. P. The mre11 complex: at the crossroads of DNA repair and checkpoint signalling. *Nature Rev. Mol. Cell Biol.* **3**, 317–327 (2002).
46. Petrini, J. H. & Stracker, T. H. The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol.* **13**, 458–462 (2003).
47. Lee, J., Kumagai, A. & Dunphy, W. G. Claspin, a Chk1-regulatory protein, monitors DNA replication on chromatin independently of RPA, ATR, and Rad17. *Mol. Cell* **11**, 329–340 (2003).
48. Sorensen, C. S., Syljuasen, R. G., Lukas, J. & Bartek, J. ATR, Claspin and the Rad9–Rad1–Hus1 complex regulate Chk1 and Cdc25A in the absence of DNA damage. *Cell Cycle* **3**, 941–945 (2004).
49. Krämer, A. *et al.* Centrosome-associated Chk1 prevents premature activation of cyclin-B-Cdk1 kinase. *Nature Cell Biol.* **9**, 884–891 (2004).
50. Bartek, J., Bartkova, J. & Lukas, J. The retinoblastoma protein pathway in cell cycle control and cancer. *Exp. Cell Res.* **237**, 1–6 (1997).
51. Sherr, C. J. & McCormick, F. The RB and p53 pathways in cancer. *Cancer Cell* **2**, 103–112 (2002).
52. Wahl, G. M. & Carr, A. M. The evolution of diverse biological responses to DNA damage: insights from yeast and p53. *Nature Cell Biol.* **3**, E277–E286 (2001).
53. Craig, A. *et al.* Allosteric effects mediate CHK2 phosphorylation of the p53 transactivation domain. *EMBO Rep.* **4**, 787–792 (2003).
54. Maya, R. *et al.* ATM-dependent phosphorylation of Mdm2 on serine 395: role in p53 activation by DNA damage. *Genes Dev.* **15**, 1067–1077 (2001).
55. Bartek, J., Lukas, C. & Lukas, J. Checking on DNA damage in S phase. *Nature Rev. Mol. Cell Biol.* **5**, 792–804 (2004).
56. Donzelli, M. & Draetta, G. F. Regulating mammalian checkpoints through Cdc25 inactivation. *EMBO Rep.* **4**, 671–677 (2003).
57. Kim, S. T., Xu, B. & Kastan, M. B. Involvement of the cohesin protein, Smc1, in Atm-dependent and independent responses to DNA damage. *Genes Dev.* **16**, 560–570 (2002).
58. Yazdi, P. T. *et al.* SMC1 is a downstream effector in the ATM/NBS1 branch of the human S-phase checkpoint. *Genes Dev.* **16**, 571–582 (2002).
59. Taniguchi, T. *et al.* Convergence of the fanconi anemia and ataxia telangiectasia signaling pathways. *Cell* **109**, 459–472 (2002).
60. Nakanishi, K. *et al.* Interaction of FANCD2 and NBS1 in the DNA damage response. *Nature Cell Biol.* **4**, 913–920 (2002).
61. Falck, J., Petrini, J. H., Williams, B. R., Lukas, J. & Bartek, J. The DNA damage-dependent intra-S phase checkpoint is regulated by parallel pathways. *Nature Genet.* **30**, 290–294 (2002).
62. Pichierri, P. & Rosselli, F. The DNA crosslink-induced S-phase checkpoint depends on ATR-CHK1 and ATR-NBS1-FANCD2 pathways. *EMBO J.* **23**, 1178–1187 (2004).
63. Costanzo, V. *et al.* An ATR- and Cdc7-dependent DNA damage checkpoint that inhibits initiation of DNA replication. *Mol. Cell* **11**, 203–213 (2003).
64. Xu, B., Kim, S.-T., Lim, D.-S. & Kastan, M. B. Two molecularly distinct G(2)/M checkpoints are induced by ionizing irradiation. *Mol. Cell Biol.* **22**, 1049–1059 (2002).
65. Nyberg, K. A., Michelson, R. J., Putnam, C. W. & Weinert, T. A. Toward maintaining the genome: DNA damage and replication checkpoints. *Annu. Rev. Genet.* **36**, 617–656 (2002).
66. Katsuhira, U., Daigo, I., Ken, S., Nobushige, N. & Noriyuki, S. CHK1, but not CHK2, inhibits Cdc25 phosphatases by a novel common mechanism. *EMBO J.* (in the press).
67. Mailand, N. *et al.* Regulation of G(2)/M events by Cdc25A through phosphorylation-dependent modulation of its stability. *EMBO J.* **21**, 5911–5920 (2002).
68. Bulavin, D. V. *et al.* Initiation of a G2/M checkpoint after ultraviolet radiation requires p38 kinase. *Nature* **411**, 102–107 (2001).
69. Xu, B., Kim, S. & Kastan, M. B. Involvement of Brca1 in S-phase and G(2)-phase checkpoints after ionizing irradiation. *Mol. Cell Biol.* **21**, 3445–3450 (2001).
70. Taylor, W. R. & Stark, G. R. Regulation of the G2/M transition by p53. *Oncogene* **20**, 1803–1815 (2001).
71. Zhou, B. B. & Bartek, J. Targeting the checkpoint kinases: chemosensitization versus chemoprotection. *Nature Rev. Cancer* **4**, 216–225 (2004).
72. Treuner, K., Helton, R. & Barlow, C. Loss of Rad52 partially rescues tumorigenesis and T-cell maturation in Atm-deficient mice. *Oncogene* **23**, 4655–4661 (2004).
73. Stewart, G. S. *et al.* The DNA double-strand break repair gene *hMRE11* is mutated in individuals with an ataxia-telangiectasia-like disorder. *Cell* **99**, 577–587 (1999).
74. Carney, J. P. *et al.* The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. *Cell* **93**, 477–486 (1998).
75. Varon, R. *et al.* Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. *Cell* **93**, 467–476 (1998).
76. O'Driscoll, M., Ruiz-Perez, V. L., Woods, C. G., Jeggo, P. A. & Goodship, J. A. A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. *Nature Genet.* **33**, 497–501 (2003).
77. Bobabilla-Morales, L. *et al.* Chromosome instability induced *in vitro* with mitomycin C in five Seckel syndrome patients. *Am. J. Med. Genet.* **123**, 148–152 (2003).
78. Fang, Y. *et al.* ATR functions as a gene dosage-dependent tumor suppressor on a mismatch repair-deficient background. *EMBO J.* **23**, 3164–3174 (2004).
79. Bassing, C. H. *et al.* Histone H2AX: a dosage-dependent suppressor of oncogenic translocations and tumors. *Cell* **114**, 359–370 (2003).
80. Celeste, A. *et al.* H2AX haploinsufficiency modifies genomic stability and tumor susceptibility. *Cell* **114**, 371–383 (2003).
81. Canman, C. E. Checkpoint mediators: relaying signals from DNA strand breaks. *Curr. Biol.* **13**, R488–R490 (2003).
82. Liu, Q. *et al.* Chk1 is an essential kinase that is regulated by Atr and required for the G(2)/M DNA damage checkpoint. *Genes Dev.* **14**, 1448–1459 (2000).
83. Lam, M. H., Liu, Q., Elledge, S. J. & Rosen, J. M. Chk1 is haploinsufficient for multiple functions critical to tumor suppression. *Cancer Cell* **6**, 45–59 (2004).
84. Hirao, A. *et al.* Chk2 is a tumor suppressor that regulates apoptosis in both an ataxia telangiectasia mutated (ATM)-dependent and an ATM-independent manner. *Mol. Cell Biol.* **22**, 6521–6532 (2002).
85. Bell, D. W. *et al.* Heterozygous germ line *hCHK2* mutations in Li-Fraumeni syndrome. *Science* **286**, 2528–2531 (1999).
86. King, M. C., Marks, J. H. & Mandell, J. B. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643–646 (2003).
87. Venkitesan, A. R. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171–182 (2002).
88. Xu, X. *et al.* Conditional mutation of Brca1 in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation. *Nature Genet.* **22**, 37–43 (1999).
89. Xu, X. *et al.* Genetic interactions between tumor suppressors Brca1 and p53 in apoptosis, cell cycle and tumorigenesis. *Nature Genet.* **28**, 266–271 (2001).
90. McPherson, J. P. *et al.* Collaboration of Brca1 and Chk2 in tumorigenesis. *Genes Dev.* **18**, 1144–1153 (2004).
91. Kraakman-van der Zwet, M. *et al.* Brca2 (XRCC11) deficiency results in radioresistant DNA synthesis and a higher frequency of spontaneous deletions. *Mol. Cell Biol.* **22**, 669–679 (2002).
92. Howlett, N. G. *et al.* Biallelic inactivation of BRCA2 in Fanconi anemia. *Science* **297**, 606–609 (2002).
93. D'Andrea, A. D. The Fanconi road to cancer. *Genes Dev.* **17**, 1933–1936 (2003).
94. Lee, H. *et al.* Mitotic checkpoint inactivation fosters transformation in cells lacking the breast cancer susceptibility gene, Brca2. *Mol. Cell* **4**, 1–10 (1999).
95. Theunissen, J. W. *et al.* Checkpoint failure and chromosomal instability without lymphomagenesis in Mre11(ATLD1/ATLD1) mice. *Mol. Cell* **12**, 1511–1523 (2003).
96. Liu, G. *et al.* Chromosome stability, in the absence of apoptosis, is critical for suppression of tumorigenesis in Trp53 mutant mice. *Nature Genet.* **36**, 63–68 (2004).

Acknowledgements We thank members of our laboratories for invaluable discussions, and apologize to colleagues whose work could only be cited indirectly. The authors are supported by the American Lebanese Syrian Associated Charities (ALSAC) of the St. Jude Children's Research Hospital and by grants from the NIH (M.B.K.), the Danish Cancer Society and the European Union (J.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Tissue repair and stem cell renewal in carcinogenesis

Philip A. Beachy^{1,4}, Sunil S. Karhadkar^{1,2} & David M. Berman^{2,3,4}

¹Department of Molecular Biology and Genetics, The Howard Hughes Medical Institute, ²Department of Pathology, ³Department of Urology and

⁴Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA (e-mail: pbeachy@jhmi.edu)

Cancer is increasingly being viewed as a stem cell disease, both in its propagation by a minority of cells with stem-cell-like properties and in its possible derivation from normal tissue stem cells. But stem cell activity is tightly controlled, raising the question of how normal regulation might be subverted in carcinogenesis. The long-known association between cancer and chronic tissue injury, and the more recently appreciated roles of Hedgehog and Wnt signalling pathways in tissue regeneration, stem cell renewal and cancer growth together suggest that carcinogenesis proceeds by misappropriating homeostatic mechanisms that govern tissue repair and stem cell self-renewal.

The tightly regulated growth of multicellular animals presents a striking contrast to single-celled organisms, which grow and divide in a manner limited only by nutrients available in the environment. The evolutionary compensation for loss of this exuberant style of growth comes in the form of organs that afford adaptability and efficiency through specialization for functions such as locomotion and reproduction, for sensing and responding to the environment, and for acquisition and use of nutrients. The assembly of such complex organs (pattern formation) requires mechanisms to establish intricate patterns of cell division and differentiation. Development of complex organs also takes longer than simple single-cell division, thus delaying the acquisition of reproductive maturity and exposing complex multicellular animals to a greater risk of tissue damage — whether from use, predation or exposure to a hostile environment.

The evolution of mechanisms that increase the complexity of animal form thus is likely to be coupled to the evolution of mechanisms for the renewal and repair of complex organs (pattern maintenance). Given this link, it is perhaps not surprising that pattern formation and pattern maintenance share common mechanisms, such as regulation by Hedgehog (Hh) and Wnt signalling pathways. These pathways play central roles in directing embryonic pattern formation, but also function post-embryonically in stem cell renewal, tissue repair and regeneration. Moreover, when aberrantly activated, these pathways can have important roles in the initiation and growth of cancer.

Here we focus on the relationship between the normal roles of Hh and Wnt pathways in pattern maintenance and on their pathological roles in the initiation and growth of malignant tumours. Using these pathways as central points of reference, we review recent developments in the area of cancer stem cells and the relationship of cancer stem cells to tissue stem cells; we concentrate in particular on stem cell renewal in the context of tissue repair as a common antecedent of cancer initiation.

Cancer, stem cells and cancer stem cells

A long-standing idea in cancer biology is that tumours arise and grow as a result of the formation of cancer stem cells, which may constitute only a minority of the cells within a tumour but are nevertheless critical for its propagation. The concept of cancer stem cells¹ dates back almost as far as the discovery of somatic stem cells in the haematopoietic system², and was firmly established experimentally in acute

myelogenous leukaemia (AML)^{3–5}. In these studies, a minority of undifferentiated cells isolated from leukaemic patients proved to be the only cells capable of reconstituting tumours on transfer into NOD/SCID (non-obese diabetic/severe combined immunodeficient) mice; the resulting tumours included a range of more differentiated cell types like those in the original leukaemia. In their cell-surface markers, in their multipotency and in their hierarchical self-renewal properties, these cancer stem cells resemble normal haematopoietic stem cells (HSCs), suggesting that the leukaemia stem cells either derive from HSCs or from more differentiated cells through acquisition of HSC properties.

Stem cells are appealing candidates as the ‘cell of origin’ for cancer because of their pre-existing capacity for self-renewal and unlimited replication^{6,7}. In addition, stem cells are relatively long-lived in comparison to other cells within tissues. They therefore have more opportunity to accumulate the multiple additional mutations that may be required to increase the rate of cell proliferation and produce clinically significant cancers. The discovery of multipotent progenitor cells with the capacity for self-renewal (that is, stem cells) outside the haematopoietic system raises the possibility that cancer stem cells could arise from other tissue stem cells and initiate other cancer types, including solid cancers. Consistent with this possibility, a defined minority of cells within many human breast cancers are the only cells able to propagate the cancer in NOD/SCID mice, resulting in the reconstitution of tumours expressing the heterogeneous surface markers that were present in the original cancer⁸. In addition, cells with some of the properties of neural stem cells (NSCs), such as the ability to produce differentiated neurons and glia *in vitro* and *in vivo* and enhanced renewal activity, have recently been isolated from brain tumours and brain-tumour-derived cell lines^{9–11}.

The general validity of the cancer stem cell concept has not been proven, as the number of cancer types for which cancer stem cells have been identified is limited. This is, however, an important issue as successful therapy depends on targeting the cells within a tumour that drive cancer growth. However, as many cancers are heterogeneous both in their cell composition and in the relative abundance of cells capable of propagating tumour growth, it will not be surprising if cancer propagation turns out to be an exclusive property of defined subsets of cells within many particular cancer types. The key to developing effective future therapies thus seems to be the identification and characterization of these cancer stem cells, and the development of drugs that specifically target them.

Another important issue for understanding the origins of cancer is the relationship between cancer stem cells and normal tissue stem cells. The best-studied cancer stem cells are those in AML, which have been isolated and individually marked before being serially transferred through several host animals^{3–5}. In most respects examined so far, these cells resemble normal HSCs, consistent with a stem cell origin for the AML-cancer stem cell. The possibility cannot be excluded, however, that cancer stem cells might be derived from more committed progenitors by genetic or epigenetic changes that confer self-renewal ability¹². Genetic or epigenetic changes that bestow and activate the ability to self-renew on a committed progenitor cell seem less likely to occur than changes that activate renewal in a stem cell — particularly as the window of cellular plasticity within which committed progenitors might acquire renewal ability is generally limited by progression towards irreversible differentiation and replicative quiescence. De-differentiation of committed progenitors back to stem cells under tightly controlled genetic conditions has nevertheless been reported in the *Drosophila* testis and ovary. For example, prematurely differentiated spermatogonia in the *Drosophila* testis can be induced to regenerate male germline stem cells on restoration of a critical signalling pathway, Jak-STAT (Janus kinase-signal transducers and activators of transcription)¹³. Moreover, cystocytes that have already begun to form oocytes within the ovary can be induced to de-differentiate to form productive female germline stem cells on ectopic activation of signalling by Dpp (Decapentaplegic), the *Drosophila* homologue of BMP4 (bone morphogenetic protein 4)¹⁴, a member of the TGF- β (transforming growth factor- β) family.

Of particular recent interest in the origin of cancer is the observation that transient Hh and Wnt pathway activities promote stem cell self-renewal in normal tissues, whereas continuous activation is associated with the initiation and growth of many types of human cancer. These pathways thus provide a potential link between the normal self-renewal of stem cells and the aberrantly regulated proliferation of cancer stem cells.

Hh and Wnt signalling in stem cell maintenance

In addition to their well-established roles in directing the patterning of embryonic tissues and structures (reviewed in refs 15–17), the Hh and Wnt pathways have more recently been implicated in the maintenance of stem or progenitor cells in a growing list of adult tissues that now include skin, blood, gut, prostate, muscle and the nervous system^{18–29} (Table 1). Evidence for a role of these pathways in stem cell maintenance functions comes from genetic interventions *in vivo* or the treatment of isolated stem cells *in vitro* (in the case of HSCs and NSCs), followed by assays for proliferation, engraftment and multilineage potential of presumptive stem cell populations. A general feature of the results of these studies is that Wnt and Hh pathway activities seem to increase presumptive stem cell number by stimulating stem cell proliferation. Thus, for example, loss of Hh signalling does not immediately obliterate hippocampal populations of neural stem cells, but affects their number by decreasing their proliferative capacity, both *in vivo* and *in vitro*^{19,20}. A similar effect of Hh pathway activity on numbers of somatic stem cells (follicle stem cells) has been noted in the *Drosophila* ovary¹⁸. Likewise, *in vitro* treatment of isolated HSCs with Wnt or Hh proteins increases their proliferative capacity and improves their ability to form colonies *in vitro* and to colonize NOD/SCID mice^{22,27}. Similarly, loss or inhibition of Wnt pathway activity in the intestine does not abrogate the initial development of normal epithelial architecture, but instead causes a progressive degradation of epithelial structure. This effect is associated with the loss of proliferative activity in the crypts, where stem cells reside^{21,26}.

In contrast to these effects of Hh and Wnt pathway activities on stem cell self-renewal, other signals within the stem cell niche seem to function more immediately in the maintenance of stem cell identity. For example, the Jak/STAT and TGF- β signalling pathways seem to specify male and female germ cell identity in the *Drosophila* testis and ovary, respectively; genetic manipulation of these pathways rapidly

and qualitatively alters cell phenotypes^{13,14}. Signals for stem cell maintenance might therefore be classified as signals with immediate effects on the maintenance of stem cell identity, and signals that regulate renewal divisions. Identity maintenance functions cannot be ruled out for Hh and Wnt pathway activities, but most evidence points to stem cell renewal as the main target in most tissues. A role of these pathways in normal stem cell renewal is consistent with their known role in the regulation of stem cell renewal genes such as *nestin* (encoding an intermediate filament protein) and *Bmi-1* (encoding a component of the Polycomb transcriptional-silencing complexes) in tumours that depend on Hh or Wnt signalling (discussed in section ‘Cancer and persistent states of repair’ below).

Hh and Wnt signalling in cancer

The roles of Hh and Wnt pathways in stem cell renewal are particularly interesting given the genetically implied connection between activity of these pathways and the initiation and growth of a substantial fraction of lethal cancers (Table 2). Familial mutations that facilitate Hh and Wnt pathway activation have been associated with increased incidence of specific brain, skin, skeletal muscle, liver and colon cancers in humans and mice, and of bladder cancer in mice. Additional studies in which pathway activities are antagonized by treatment with pharmacological agents, with antibodies that bind and block ligand action, or by overexpression of negatively acting pathway components further demonstrate an ongoing requirement for pathway activity in the growth of additional cancer types which include small-cell lung cancer and carcinomas of the oesophagus, stomach, pancreas, biliary tract and prostate. The range of organs from which Hh- and Wnt-pathway-dependent cancers originate is therefore similar to the range of organs in which these pathways have a role in stem cell renewal. In terms of medical significance, about one-third of total cancer deaths are caused by the cancer types in which current evidence implicates Hh or Wnt pathway activity in most cases³⁰.

The Hh pathway in cancer

The link between Hh pathway activity and cancer was initially established by the identification of heterozygous mutations affecting Patched (PTCH), a negatively acting component of the Hh receptor, as the cause of Gorlin’s syndrome. This syndrome is associated with an increased incidence of basal cell carcinoma, medulloblastoma, and rhabdomyosarcoma, and PTCH is also mutated in sporadic forms of

Table 1 Patterning pathways and stem cell maintenance

Pathway	Tissue	Evidence	References
Hh	NSC	Hh required for NSC proliferation in neurospheres; conditional inactivation of SMO inhibits NSC proliferation <i>in vitro</i> and <i>in vivo</i>	19, 20
	HSC	<i>In vitro</i> and <i>in vivo</i> expansion of HSCs by Shh	27
	<i>Drosophila</i> ovary	Proliferation of ovarian somatic stem cells requires Hh	18
Wnt	HSC	Decreased proliferation and colony-forming ability of HSCs subjected to Wnt pathway inhibition	22
		Increased number of haematopoietic progenitors from mouse fetal livers and human bone-marrow cells stimulated with Wnt ligand	84, 85
	Intestine	TCF4-null mice have diminished intestinal stem cell pool	21, 26 and refs therein
	Muscle	Culture with Wnt ligand induces myogenic stem cells; Wnt antagonism with secreted Fz proteins reduces muscle stem cell proliferation	23
	Mammary gland	Increased progenitors in transgenic mice with activated Wnt signalling	55

TCF4, a TCF family member (see text; Fig. 1).

these cancers, thus identifying PTCH as a tumour suppressor (reviewed in refs 6, 31). A similar range of tumours was also found to be associated with sporadic activating mutations affecting the positive receptor component and proto-oncogene Smoothened (SMO). In normal Hh pathway function, the transporter-like Ptch protein acts catalytically to restrain activation of the seven-transmembrane protein Smo. Ptch activity is blocked by binding of Hh, which liberates Smo for activation of transcriptional targets through the Gli family of latent transcriptional factors (see Fig. 1a). These features of Hh signalling are broadly conserved between *Drosophila* and mammals, as are common mechanisms for Hh protein processing and lipid modification and a dedicated mechanism for the release of lipid-modified Hh protein from producing cells. (reviewed in refs 32, 33). In addition the Gli proteins, like their *Drosophila* counterpart Ci, can be regulated by interactions with the Suppressor of fused (Su(fu)) protein and can exist in activating forms (primarily Gli and Gli2) as well as in proteolytically processed repressing forms (primarily Gli3). The human SU(FU) gene has also been implicated as a tumour suppressor, with mutations found in familial and sporadic medulloblastoma and in sporadic basal cell carcinoma (see Table 2).

Despite extensive similarities between *Drosophila* and mammalian pathways, however, significant differences may exist, particularly in the transducing machinery between Smo and Gli. Thus, although recent genetic and biochemical studies in *Drosophila* have demonstrated that pathway activation is transmitted through association of Smo with a complex of cytoplasmic proteins that includes Ci and a kinesin-like protein, Cos2, a functional mammalian homologue of Cos2 has not been identified (reviewed in ref. 32). Because of its role in maintenance of pathway quiescence in *Drosophila*, a functional mammalian Cos2 homologue would be of interest as a potential tumour suppressor. In addition, several apparent pathway components identified in mammals either have no counterparts or do not function in the *Drosophila* Hh pathway (see ref. 34). These include components such as RAB23 (ref. 35) or FKBP8 (ref. 34), which have unknown function, but are of interest as potential tumour suppressors because of their action downstream of Smo as negative regulators of pathway activity (see Fig. 1).

Some tumours of the type associated with Gorlin's syndrome are not associated with known pathway-activating mutations, despite clear evidence for pathway activity^{36,37}. This suggests that activation of the Hh pathway may occur through mechanisms other than by mutation of pathway components, and raises the possibility that such mechanisms may also have a role in pathway activation in other cancers not typically associated with Gorlin's syndrome. Consistent with this possibility, recent studies using cyclopamine, a specific Hh pathway antagonist^{38–40}, indeed have demonstrated an ongoing requirement for pathway activity in the growth of a series of lethal cancers arising in organs of endodermal origin, and not typically associated with Gorlin's syndrome. These cancers include small-cell lung cancer and carcinomas of the oesophagus, stomach, pancreas, biliary tract, and prostate^{41–43}. Pathway activity in these cancers requires ligand activation, as demonstrated with the use of Hh-blocking antibodies, and contrary to ligand-independent activation arising in tumours associated with Gorlin's syndrome. Curiously, the limiting factor in pathway activation in these non-Gorlin's tumours seems not to be ligand expression, but rather the acquisition of responsiveness to ligand. Thus, whereas the Hh family members Shh (Sonic hedgehog) and Ihh (Indian hedgehog) are expressed in normal endodermal tissues, high-level activation of Hh pathway targets occurs only in cancer cells. In the prostate, the limiting factor for ligand responsiveness is SMO, which is not expressed in normal prostate tissue²⁹. Furthermore, isolated prostate stem/progenitor cells acquire Hh responsiveness simply by introduction of Smo expression constructs, and these cells are oncogenically transformed upon pathway activation. The genetic or epigenetic changes that trigger Smo expression are not identified, although they may be linked to epithelial regeneration (see section 'Cancer and persistent states of repair' below).

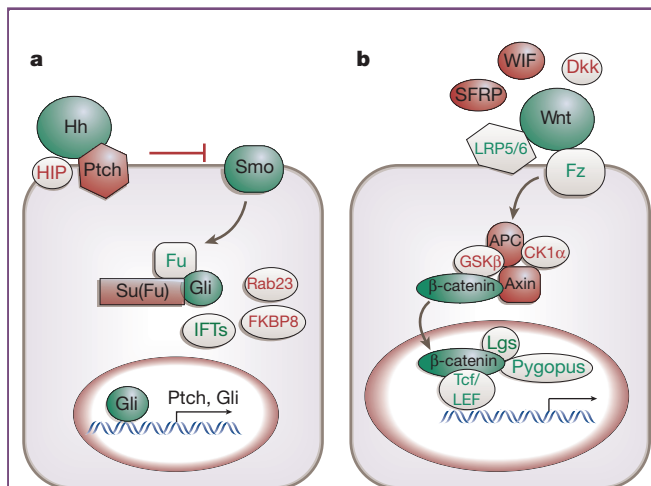


Figure 1 Hh and Wnt signalling pathways. Simplified views of the Hh and Wnt signalling pathways, with emphasis on components implicated in cancer or tissue regeneration. Green and red colours denote pathway components with primarily positive or negative roles, respectively, in pathway activation. Shaded components have been causally implicated in tumorigenesis (see Table 2 and text; more complete pathway descriptions are available in refs 32–34 for Hh and refs 17, 46 for Wnt). **a**, Activation of the Hh signalling pathway is initiated by binding of a Hh ligand to Ptch. This lifts suppression of Smo, activating a cascade that leads to the nuclear translocation of Gli and the activation of target genes. HIP is a membrane protein that antagonizes pathway activity by binding to Hh ligands, and Fu, Su(Fu), Rab23, FKBP8 and the IFTs (intraflagellar transport proteins) act downstream of Ptch and Smo to regulate Gli. The function of Rab23, FKBP8 and the IFTs outside the CNS is not established. HIP, Hh-interacting protein; Rab23, a member of the Rab family of GTPases; FKBP8, a member of the FK506-binding protein family. **b**, The Wnt signalling pathway is activated by binding of Wnt ligands to their receptors Fz and LRP5/6, leading to the release of β -catenin from the degradation complex and facilitating its entry into the nucleus, where it regulates target gene transcription through association with Tcf/LEF, Legless (Lgs) and Pygopus. SFRP, WIF and Dkk are secreted antagonists of Wnt signalling. APC, Axin, GSK3 β and CK1 α are components of the β -catenin degradation complex. WIF, Wnt inhibitory factor; Dkk, Dickkopf; GSK3 β , glycogen synthase kinase 3 β ; CK1 α , casein kinase 1 α .

The Wnt pathway in cancer

The Wnt signalling pathway has also been implicated in several types of cancer, initially through overexpression of the Wnt-1 protein signal in murine mammary tumours as a consequence of nearby mouse mammary tumour virus (MMTV) insertion⁴⁴, and subsequently through the substantially increased incidence of colorectal and other cancers in familial adenomatous polyposis, caused by mutations affecting the tumour suppressor APC (adenomatous polyposis coli) (reviewed in ref. 45). In the absence of Wnt signal, APC fosters the degradation of the oncogene β -catenin and prevents its entry into the nucleus (Fig. 1b). Wnt stimulation, loss of APC protein function, or of its associated partner Axin, all lead to the stabilization of β -catenin and to its increased concentration in the nucleus. β -catenin can then act as a transcriptional co-activator by associating with the Tcf/LEF family of transcription factors. A complex of APC with Axin and other proteins targets β -catenin for proteasomal degradation by scaffolding an association between β -catenin and kinases whose activities lead to β -catenin ubiquitinylation. This action is abrogated by the recruitment of the degradation complex to the membrane upon Wnt activation of a receptor complex that includes Frizzled (Fz), a relative of Smo, and LRP5/6. Any lesion causing β -catenin accumulation through the disruption of a degradation complex component or by mimicking complex recruitment to the receptor would be expected to promote tumour formation. This pathway can also be activated by mutations of β -catenin that render it resistant to degradation (for detailed reviews of Wnt signalling see refs 6, 17, 46).

Table 2 Hh and Wnt pathways in cancer

Tissue	Tumour	Evidence of pathway involvement	References
Hh pathway			
Brain	Medulloblastoma	Tumorigenesis by inactivation of <i>PTCH</i> ; allograft and cell-line growth inhibition by cyclopamine; inhibition of autochthonous tumour growth by synthetic small molecule antagonist	37, 81; reviewed in 6
		Tumorigenesis by inactivation of <i>Su(fu)</i>	86
	Glioma	Gli amplification; growth inhibition of some cell lines by cyclopamine	87, 88
Skin	Basal cell carcinoma	Tumorigenesis by inactivation of <i>PTCH</i> ; <i>in vivo</i> tumorigenesis by expression of activating form of <i>SMO</i> or by Shh overexpression and <i>in vitro</i> growth inhibition by synthetic Hh pathway antagonist; inhibition of human tumour growth topical cyclopamine	82, 83; reviewed in 6
Muscle	Rhabdomyosarcoma	Tumorigenesis by inactivation of <i>PTCH</i>	reviewed in 6
Oesophagus	Adenocarcinoma	Cell-line growth inhibition by cyclopamine, Hh blocking antibody	42
Stomach	Adenocarcinoma	Cell-line growth inhibition by cyclopamine, Hh blocking antibody	42
Pancreas	Adenocarcinoma	Xenograft and cell-line growth inhibition by cyclopamine, Hh blocking antibody; tumour initiation (in mouse) by Shh overexpression	42, 43
Biliary tract	Adenocarcinoma	Xenograft and cell-line growth inhibition by cyclopamine, Hh blocking antibody	42
Lung	Small-cell lung cancer	Xenograft and cell-line growth inhibition by cyclopamine, Hh blocking antibody	41
Prostate	Adenocarcinoma	Xenograft and cell-line growth inhibition and suppression of metastasis by cyclopamine; increased xenograft growth by Shh and Gli overexpression	29, 89, 90
Bladder	Urothelial carcinoma	Increased tumour induction (in mouse) by alkylating agent in <i>Ptch</i> heterozygote	91
Oral cavity	Squamous cell cancer	Growth inhibition of cell lines by cyclopamine;	92
Wnt pathway			
Colon	Adenocarcinoma	Tumorigenesis by inactivation of APC, Axin; tumorigenesis by stabilization of β -catenin; epigenetic inactivation of SFRPs	47; reviewed in 45
Liver	Hepatoblastoma	Tumorigenesis (in mouse) by inactivation of APC and by stabilization of β -catenin	reviewed in 45
Blood	Multiple myeloma	Cell-growth inhibition by dominant negative TCF4; growth stimulation by Wnt ligand	93
Hair follicle	Pilomatricoma	Tumorigenesis (in mouse) by overexpression of β -catenin	reviewed in 45
Bone	Osteosarcoma	Dkk3 and LRP5 expression inhibits tumour cell growth <i>in vitro</i>	94, 95
Lung	Non-small-cell carcinoma	Apoptosis and cell-growth inhibition by short interfering RNA and a blocking antibody against Wnt2	96
Pleura	Mesothelioma	Apoptosis and cell-growth inhibition by transfection of SFRP	97

Emphasis is placed on functional data showing a requirement for pathway activation in tumour formation and/or tumour cell growth. (See Fig. 1 and text for gene abbreviations.)

Hh and Wnt signalling pathways are similar in that both signals are lipidated (an important process that affects their activity and tissue distribution³³), and they use several related or identical components. The fundamental logic of pathway activation is also similar, in both cases involving receptor recruitment of multicomponent complexes with key roles in cytoplasmic anchoring and proteolysis of key transcriptional effectors. It is also possible that, like Hh, the Wnt pathway is activated in a wider range of cancers than has been revealed by familial or sporadic mutations that produce ligand-independent pathway activation. The possibility of Wnt ligand dependent pathway activation in cancer is suggested by a recent study demonstrating that epigenetic silencing of *SFRPs* (secreted frizzled-related proteins), which encode an extracellular ligand-binding pathway antagonist, may have a critical role in the early establishment of colorectal cancer⁴⁷. A broader range of cancers requiring Wnt pathway activity for growth may also be revealed, as potent and specific Wnt pathway antagonists are identified and become broadly available⁴⁸.

Hh and Wnt pathways in regeneration and tissue repair

If cancer stem cells arise from tissue stem cells, and if Hh and Wnt pathway activities are critical for the renewal of at least some of these stem cell types, then continuous Hh and Wnt pathway activities may promote cancer growth by continuously recapitulating their roles in promoting normal stem cell renewal. But stem cell renewal must be tightly regulated (otherwise tumours might arise), raising the critical question of how and under what circumstances normal regulation can be circumvented in cancer.

Some insight into the regulation of stem cell renewal activity may be gained from a consideration of the role of Hh and Wnt pathways in regenerative responses (Table 3). Wnt pathway activation in the radially symmetric coelenterate *Hydra* is closely associated with the growth and patterning of new individuals. This may result either

from normal asexual budding or from experimental manipulations, such as cell dissociation and re-aggregation⁴⁹. *Hydra* tissue thus seems to exist in a constant state of growth and replacement. Amphibia, particularly urodeles (newts and salamanders), are also capable of mounting impressive regenerative responses to limb amputation or to extirpation of certain organs. The typical sequence of events involves de-differentiation of cells near the site of injury, followed by extensive proliferation and patterning of the regenerating tissues. In the cases of urodele limb and lens regeneration, Hh family members are expressed in the de-differentiated cells following injury, and regeneration can be blocked by treatment with cyclopamine^{50–52}. Fin regeneration in fish also entails expression of Hh genes and targets, and is disrupted by cyclopamine inhibition^{53,54}.

The regenerating structures in these examples encompass several tissue types that are arranged in complex patterns; in this respect pathway roles resemble those in embryonic pattern formation. But Hh and Wnt pathway activity also have a role in regenerative responses that are restricted to single tissue types within organs. For example, transient Hh pathway activity is required for androgen-triggered regeneration of prostate epithelium in male mouse castrates²⁹, and Wnt pathway activities similarly are required for muscle regeneration in response to cardiotoxin-induced muscle injury²³. Increased Hh pathway activity in *Ptch*^{+/–} mice also contributes to an increase in photoreceptor-cell progenitor number and retinal repair in a model of retinal degeneration. Furthermore, mammary progenitors are enriched in mice with Wnt pathway activation caused by increased Wnt ligand levels or by a β -catenin altered to increase its stability⁵⁵. In addition to these demonstrations of functional Hh and Wnt pathway activity in tissue repair, correlative data suggest a possible role for Wnt pathway activity in response to biliary tract⁵⁶, liver⁵⁷ and kidney tubule injury⁵⁸, and for Hh pathway activity in repair of bone fractures⁵⁹, bile duct⁵⁶ and airway injury⁴¹.

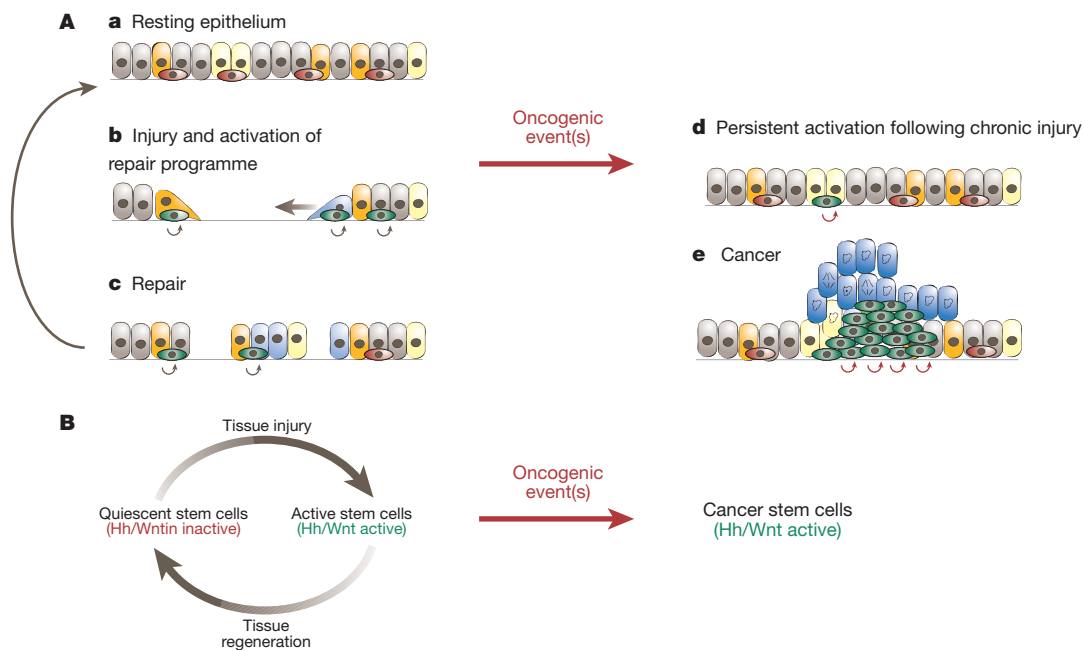


Figure 2 Model for carcinogenesis resulting from persistence of a state of injury repair.

A. Cellular events of epithelial repair. **a.** Resting epithelium with several differentiated cell phenotypes (brown, orange, and yellow) derived from tissue stem cells, now quiescent (red). Pathways such as Hh and Wnt signalling pathways that have a role in the renewal of stem cells are not active. **b.** Epithelial defect resulting from acute injury. Loss of epithelial continuity activates a repair program which is driven by Hh or Wnt signalling. This program results in the acquisition by epithelial cells of a more mesenchymal phenotype, including flattening and movement of cells (straight arrow) to cover the wound, activation (green), and expansion of stem cells through renewal divisions (curved arrows). **c.** The wound is repaired, first by rapid cell movement, and then by restoration of cell numbers resulting from the amplification of stem cells and

the differentiation of their progeny. Subsequently, either epithelial continuity and patterning is restored, Hh and Wnt signalling ceases, and the stem cell compartment returns to quiescence (**a**); or oncogenic event(s) may trap a stem cell in an activated state of continuous renewal, which is driven by autonomous Wnt or Hh signalling (**d**). Further genetic or epigenetic change in such a persistently activated stem cell (curved red arrows) might produce a cancer stem cell (green) which is capable of aggressively propagating a cancer (**e**). This may result from enhanced proliferation and production of more cancer stem cells as well as from differentiated cancer cells (blue). **B.** Stem cells cycle between quiescence and activity as a consequence of Hh/Wnt driven responses to injury. Oncogenic event(s) may trap activated stem cells in a permanent state of Hh/Wnt driven activity, resulting in cancer stem cells.

Cancer and persistent states of repair

We have reviewed evidence highlighting the role of Hh and Wnt pathway activity in cancer growth on the one hand, and in stem cell renewal and tissue regeneration on the other. But is there a link between tissue repair and cancer? A connection is strongly suggested by the known association between chronic tissue injury and cancer^{60,61}, including cancers associated with Hh and/or Wnt pathway activity. Increased cancer risks are associated with exposure to toxins, such as alcohol, cigarette smoke and organic chemicals^{62–64}, with chronic infection of *Helicobacter pylori* and other pathogens⁶⁵, and with inflammatory conditions, such as sclerosing cholangitis and inflammatory bowel disease^{66,67} — all of which entail chronic tissue damage. As discussed above, acute injury is accompanied by the expansion of stem cell pools and by the transient activation of the Hh and Wnt signalling pathways⁴¹. Under conditions of chronic injury, pathway activation and presumed expansion of stem cell pools would persist so long as repeated injury prevents full regeneration. This state of continuous pathway and progenitor-cell activation resembles the continuous pathway activity and cell division seen in cancer.

These observations suggest that cancer growth may represent the continuous operation of an unregulated state of tissue repair and that continuous Hh/Wnt pathway activities in carcinogenesis may represent a deviation from the return to quiescence that normally follows regeneration (Fig. 2A, a, b). The simplest model for the emergence of this state is that genetic or epigenetic events prevent the return to quiescence of an activated stem or progenitor cell on completion of regeneration, thus initiating a tumour by trapping the cell in an activated state of continuous renewal. Consistent with this model,

the *Bmi-1* gene required for HSC renewal is also required for the propagation of leukaemias in transfer experiments^{68,69}. The expression of *Bmi-1* and nestin, which are both associated with stem cell renewal^{68–70}, is dependent on Hh pathway activity in Hh-dependent tumours^{29,37,41,71}. Of course, multiple genetic or epigenetic changes might be required to trap the activated stem cell initially, and numerous other events may contribute to rapid proliferation or to other aspects of the phenotype. Conversion of an activated stem cell into a clinically threatening cancer stem cell may involve changes that lock the cell in an active state of renewal and allow the cell to acquire independence from niche signals that are normally required to maintain stem cell identity.

The observed increase in cancer incidence associated with chronic injury strongly supports this model of cancer as a continuous state of repair. If, as hypothesized, the oncogenic event results in trapping activated stem cells in a continual state of renewal, then any condition that increases the pool size of activated stem cells should increase the probability of an oncogenic event by making the cellular substrates for such an event more numerous. The effect of repeated injury over time would be exactly this — to increase the pool size of stem cells in an active state of renewal (Fig. 3). Tissues that normally undergo rapid renewal might also be expected to experience increased cancer incidence, as a high turnover rate might require a sizeable pool of activated stem cells. Indeed, organs such as the skin, the lungs, and the gastrointestinal tract, which are continuously exposed to environmental insults, and consequently in a constant state of renewal, are the tissues of origin for a high proportion of cancers.

Table 3 **Hedgehog and Wnt signalling in regeneration**

Organism	Organ/tissue	Injury	Components induced during regeneration	Pathway modulation and effect	References
Hh pathway					
Newt	Lens	Lensectomy	Shh, Ihh, Ptch	Regeneration blocked by *KAAD-cyclopamine and HLP-transfection	50
	Limb	Amputation	Shh, Ptch	Regeneration blocked by cyclopamine	51, 52
Zebrafish	Fin	Amputation	Shh, Ptch, Bmp2	Regeneration blocked by cyclopamine	53, 54
Mouse	Vasculature	Ischaemia	Shh, Dhh, Ptch	Regeneration blocked by anti-Hh antibody	98
	Prostate	Androgen deprivation by castration	Ptch	Regeneration blocked by cyclopamine and anti-Hh antibody	29
	Retina	Degeneration	Ptch	Mice with Ptch mutations show improved retinal photoreceptor regeneration	99
	Facial nerve	Axotomy	Shh, Smo	Regeneration-reduced by cyclopamine	100
	Bile duct	Immune	Ptch, Smo		56
	Lung	Chemical	Ptch, Gli		41
	Bone	Fracture	Shh, Ptch, Smo, Gli		59, 72
Wnt pathway					
Mouse	Muscle	Ischaemia	Wnts 5a, 5b, 7a, 7b, β -catenin, Tcf	Muscle regeneration inhibited by secreted Fz peptides (SFRPs)	23
Human	Bile duct	Immune	Wnts 2, 5a, 10b, 12, 13		56
Human	Kidney	Ureteral obstruction	Wnt 4		58
Hydra	Axis	Dissociation	Tcf, Wnt, β -catenin		49
Mouse	Liver	Hepatectomy	Wnt1, β -catenin		57

Several of the organs listed also require Hh and Wnt signalling pathways during embryonic patterning. Where available, experimental evidence for pathway function in regeneration is listed under 'Pathway modulation and effect'. * A more potent derivative of cyclopamine.

The nature of the oncogenic events that may trap stem cells in an active state of renewal is not always clear. As noted above, Hh pathway activation in tissues that give rise to non-Gorlin's tumours seem not to be limited by ligand availability, but by the responsiveness to ligand. In normal prostate, the limiting factor in pathway responsiveness is SMO expression, and SMO upregulation is uniformly noted in all metastatic prostate cancer²⁹. Upregulation of Smo also occurs in mice in response to injury of other endodermal tissues (P.A.B., S.S.K. and D.M.B., unpublished data), and has been dramatically demonstrated to occur locally at the site of bone injury⁷². The acquisition of pathway responsiveness through SMO upregulation is therefore a common feature of both injury response and tumorigenesis; cancer cells in this respect closely resemble cells in injured tissues. The identity and source of the signal that induces Smo expression in injured tissues may therefore lend insight into the targets of oncogenic processes that lead to SMO expression and Hh ligand responsiveness.

Tissue repair, invasion and metastasis

Most cancer deaths are caused by the growth of tumours at distant metastatic sites rather than at the site of origin. Metastasis requires the capacity to detach from the original tumour mass, to migrate through several types of tissue and to colonize a permissive ectopic site. Current evidence suggests that there may be associations between the activities of Hh and Wnt pathways and metastasis, at least in some types of cancer. In colorectal and pancreatic adenocarcinomas, Wnt and Hh pathway activities, respectively, have been linked to all stages of these diseases — from pre-invasive neoplasia to locally growing and metastatic lesions^{42,43,73}.

In contrast, high-level Hh pathway activity in prostate cancer is associated exclusively with metastatic tumours²⁹, and cell lines with and without the ability to metastasize can be interconverted by modulation of Hh pathway activity. Hh signalling specifically promotes collagen-matrix invasion and the expression of genes associated with a transition from epithelial to mesenchymal character (epithelial-mesenchymal transition, EMT). These genes include

Snail, a helix-loop-helix transcriptional repressor⁷⁴. EMT and the expression of Snail or its homologue Slug is also associated with aggressive behaviour, including metastasis, of other cancers^{74,75}, and has been linked to activity of the Wnt pathway in colorectal cancer^{76,77}.

How does the promotion of invasiveness seen in tumours relate to the physiological roles of Hh or Wnt? Migration through tissues is a normal feature of neural crest, germ cell and haematopoietic stem cell development, and is also observed during acute epithelial injury repair in adults, in a process called epithelial restitution⁷⁸. During restitution, epithelial cells adjacent to a focally denuding injury detach from each other, assume an elongated shape and rapidly migrate (often within 2–4 hours) to the injured area, where they invade remnants of the injured tissue and reconstitute epithelial continuity. This behaviour has so far been observed in differentiated

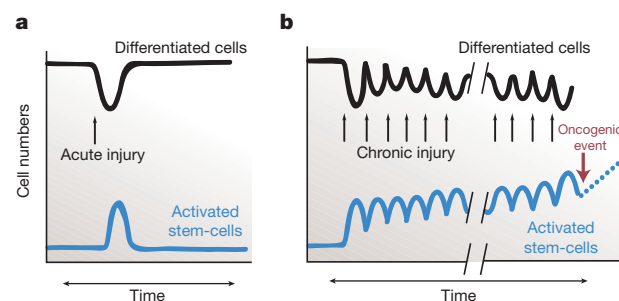


Figure 3 Increased cancer risk during chronic injury. **a**, Acute injury produces a transient expansion in the pool of activated stem cells. On completion of repair, the number of activated stem cells returns to baseline. **b**, Chronic injury also causes expansion in the pool of activated stem cells, but because of repeated injury the pool size does not return to baseline. As a result, the total number of activated stem cells is greatly increased over time, increasing the probability that an oncogenic event will trap a stem cell in the activated state.

cells, but as convenient markers for tracking stem cells in such experiments are lacking, the possibility remains that stem cells also become motile and invasive during injury repair. The acquisition of such an ability to invade and move through tissues might represent part of a programme of stem cell activation for optimal completion of epithelial repair. Hh and Wnt pathway activities are thus linked to the activation of stem cells in injury repair, and this reparative state is associated with cell behaviour that is recapitulated in metastasis. It seems plausible therefore that the trapping of stem cells in a state of repair might predispose them not only to tumour formation in general, but to the formation of aggressive tumours in particular.

Perspectives and implications

If cancer stem cells responsible for driving the growth of cancer types associated with Hh and Wnt pathway activation indeed come from stem cells trapped in a state of active renewal by pathway activities, then a logical therapeutic approach for these cancers would be to impose a state of pathway blockade (see introduction in this issue by Sawyers, page 294). Potential problems associated with such approaches might include cognitive or affective disturbances, as both Hh and Wnt pathways are active in the mature brain. In addition, the roles of pathway activities in normal stem cell renewal suggest that pathway blockade might cause a complete or partial loss of stem cell pools. Latent symptoms caused by such a loss might include an increased susceptibility to degenerative disorders, and appear only after passage of a significant fraction of lifespan. On the other hand, Hh or Wnt pathway activities might be restricted solely to the stimulation of stem cell self-renewal and not affect signals required for the maintenance of stem cell identity. In this case, endogenous stem cells may remain quiescent during pathway blockade but regain renewal capacity once therapy is completed and the blockade lifted. Stem cell niches might also persist and permit the regeneration of stem cells that are temporarily lost during a period of pathway blockade. Consistent with such a possibility, the well-characterized germline stem cell niches in the *Drosophila* ovary and testis have been reported to persist and supply continuous niche signals for a significant fraction of the adult lifespan, even after they are emptied of stem cells^{13,79}.

The potential success of such therapeutic approaches is suggested by the achievement of growth inhibition or regression, complete in some cases with systemic treatment by the Hh pathway antagonist cyclopamine in murine models of several Hh-pathway-associated tumour types^{29,37,42,43,80}. In addition, a recent report demonstrates growth inhibition of spontaneous medulloblastomas arising in *Ptch*^{+/-}*p53*^{-/-} mice on systemic treatment with a synthetic Hh pathway antagonist⁸¹. Cancer growth in these tumour types apparently requires an active state of renewal, without which cancer stem cells are depleted by differentiation or apoptosis. The lack of toxic effects in mice during periods of systemic cyclopamine treatment extending as long as seven weeks and during follow-up observation periods of nearly half a year also augurs well for this approach. More recently, cyclopamine-induced regression of human basal-cell carcinomas was reported⁸², suggesting the potential effectiveness of Hh pathway blockade in humans. As cyclopamine application in these human cases was topical, cognitive or affective disturbances that might be caused by systemic pathway blockade cannot be ruled out. Such effects, if they materialize, might be reduced or eliminated by the development of pathway-blocking agents that do not cross the blood/brain barrier. The feasibility of such an approach is suggested by the identification of several structurally distinct classes of Hh pathway antagonists^{40,83}. Some recent success has also been reported in the identification of Wnt pathway antagonists⁴⁸, suggesting that the therapeutic effects of blocking this pathway may be tested in the near future.

Systemic pathway blockade in humans may require consideration of other factors. For example, the inability of prostate epithelium or muscle to regenerate under conditions of Hh or Wnt pathway blockade may be typical of many tissues, and the loss of stem cell renewal divisions could result in increased sensitivity to injury or

other transient demands being placed on stem cell pools. It may therefore be important for patients to avoid even mild sources of trauma while undergoing pathway blockade therapy. This consideration also raises doubts about combining any form of cytotoxic chemotherapy with pathway blockade, as indiscriminate injury imposed by such therapy might affect some of the tissues containing stem cells whose renewal depends on pathway activities. Despite these concerns, preliminary studies *in vitro* and in mice suggest that blockade of these pathways may offer a new and efficacious therapeutic approach to a large group of highly lethal cancers. Other strategies of potential use in cancer prevention and therapy might also arise from an improved understanding of the response to tissue injury, in particular of the signals that regulate the activation of tissue stem cells. □

doi:10.1038/nature03100

- Park, C. H., Bergsagel, D. E. & McCulloch, E. A. Mouse myeloma tumour stem cells: a primary cell culture assay. *J. Natl Cancer Inst.* **46**, 411–422 (1971).
- Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14**, 213–222 (1961).
- Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
- Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3**, 730–737 (1997).
- Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunol.* **5**, 738–743 (2004).
- Taipale, J. & Beachy, P. A. The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001).
- Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
- Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
- Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63**, 5821–5828 (2003).
- Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumors. *Proc. Natl Acad. Sci. USA* **100**, 15178–15183 (2003).
- Kondo, T., Setoguchi, T. & Taga, T. Persistence of a small subpopulation of cancer stem-like cells in the C6 glioma cell line. *Proc. Natl Acad. Sci. USA* **101**, 781–786 (2004).
- Passague, E., Jamieson, C. H., Ailles, L. E. & Weissman, I. L. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? *Proc. Natl Acad. Sci. USA* **100** (suppl. 1), 11842–11849 (2003).
- Brawley, C. & Matusis, E. Regeneration of male germline stem cells by spermatogonial dedifferentiation *in vivo*. *Science* **304**, 1331–1334 (2004).
- Kai, T. & Spradling, A. Differentiating germ cells can revert into functional stem cells in *Drosophila melanogaster* ovaries. *Nature* **428**, 564–569 (2004).
- Ingham, P. W. & McMahon, A. P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev.* **15**, 3059–3087 (2001).
- Muenke, M. & Beachy, P. A. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C. *et al.*) 6203–6230 (McGraw-Hill, New York, 2001).
- Logan, C. Y. & Nusse, R. The Wnt signaling pathway in development and disease. *Annu. Rev. Cell. Dev. Biol.* **20**, 781–810 (2004).
- Zhang, Y. & Kalderon, D. Hedgehog acts as a somatic stem cell factor in the *Drosophila* ovary. *Nature* **410**, 599–604 (2001).
- Machold, R. *et al.* Sonic hedgehog is required for progenitor cell maintenance in telencephalic stem cell niches. *Neuron* **40**, 189–190 (2003).
- Lai, K., Kaspar, B. K., Gage, F. H. & Schaffer, D. V. Sonic hedgehog regulates adult neural progenitor proliferation *in vitro* and *in vivo*. *Nature Neurosci.* **6**, 21–27 (2003).
- Pinto, D., Gregorieff, A., Begthel, H. & Clevers, H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes Dev.* **17**, 1709–1713 (2003).
- Reya, T. *et al.* A role for Wnt signalling in self-renewal of haematopoietic stem cells. *Nature* **423**, 409–414 (2003).
- Poleskaya, A., Seale, P. & Rudnicki, M. A. Wnt signaling induces the myogenic specification of resident CD45⁺ adult stem cells during muscle regeneration. *Cell* **113**, 841–852 (2003).
- Owens, D. M. & Watt, F. M. Contribution of stem cells and differentiated cells to epidermal tumours. *Nature Rev. Cancer* **3**, 444–451 (2003).
- Perez-Losada, J. & Balmain, A. Stem-cell hierarchy in skin cancer. *Nature Rev. Cancer* **3**, 434–443 (2003).
- Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature Genet.* **19**, 379–383 (1998).
- Bhardwaj, G. *et al.* Sonic hedgehog induces the proliferation of primitive human hematopoietic cells via BMP regulation. *Nature Immunol.* **2**, 172–180 (2001).
- Ramalho-Santos, M., Melton, D. A. & McMahon, A. P. Hedgehog signals regulate multiple aspects of gastrointestinal development. *Development* **127**, 2763–2772 (2000).
- Karhadkar, S. S. *et al.* Hedgehog signalling in prostate regeneration, neoplasia and metastasis. *Nature* **431**, 707–712 (2004).
- Cancer facts and figures. *Am. Cancer Soc.* <http://www.cancer.org/> (2003).
- Wechsler-Reya, R. & Scott, M. P. The developmental biology of brain tumors. *Annu. Rev. Neurosci.* **24**, 385–428 (2001).
- Lum, L. & Beachy, P. A. The Hedgehog response network: sensors, switches, and routers. *Science* **304**, 1755–1759 (2004).

33. Mann, R. K. & Beachy, P. A. Novel lipid modification of secreted protein signals. *Annu. Rev. Biochem.* **73**, 891–923 (2004).
34. Bulgakov, O. V., Eggenschwiler, J. T., Hong, D. H., Anderson, K. V. & Li, T. F. K. Bmp8 is a negative regulator of mouse sonic hedgehog signaling in neural tissues. *Development* **131**, 2149–2159 (2004).
35. Eggenschwiler, J. T., Espinoza, E. & Anderson, K. V. Bmp23 is an essential negative regulator of the mouse Sonic hedgehog signalling pathway. *Nature* **412**, 194–198 (2001).
36. Dahmane, N., Lee, J., Robins, P., Heller, P. & Altaba, R. I. Activation of the transcription factor Gli1 and the Sonic hedgehog signalling pathway in skin tumours. *Nature* **389**, 876–881 (1997).
37. Berman, D. M. *et al.* Medulloblastoma growth inhibition by hedgehog pathway blockade. *Science* **297**, 1559–1561 (2002).
38. Taipale, J. *et al.* Effects of oncogenic mutations in Smoothened and Patched can be reversed by cyclopamine. *Nature* **406**, 1005–1009 (2000).
39. Chen, J. K., Taipale, J., Cooper, M. K. & Beachy, P. A. Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened. *Genes Dev.* **16**, 2743–2748 (2002).
40. Chen, J. K., Taipale, J., Young, K. E., Maiti, T. & Beachy, P. A. Small molecule modulation of Smoothened activity. *Proc. Natl Acad. Sci. USA* **99**, 14071–1406 (2002).
41. Watkins, D. N. *et al.* Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer. *Nature* **422**, 313–317 (2003).
42. Berman, D. M. *et al.* Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours. *Nature* **425**, 846–851 (2003).
43. Thayer, S. P. *et al.* Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. *Nature* **425**, 851–856 (2003).
44. Nusse, R. & Varmus, H. E. Many tumors induced by the mouse mammary tumour virus contain a provirus integrated in the same region of the host genome. *Cell* **31**, 99–109 (1982).
45. Giles, R. H., van Es, J. H. & Clevers, H. Caught up in a Wnt storm: Wnt signaling in cancer. *Biochim. Biophys. Acta* **1653**, 1–24 (2003).
46. He, X., Semenov, M., Tamai, K. & Zeng, X. LDL receptor-related proteins 5 and 6 in Wnt/beta-catenin signaling: arrows point the way. *Development* **131**, 1663–1677 (2004).
47. Suzuki, H. *et al.* Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nature Genet.* **36**, 417–422 (2004).
48. Lepourclet, M. *et al.* Small-molecule antagonists of the oncogenic Tcf/beta-catenin protein complex. *Cancer Cell* **5**, 91–102 (2004).
49. Hobmayer, B. *et al.* WNT signalling molecules act in axis formation in the diploblastic metazoan Hydra. *Nature* **407**, 186–189 (2000).
50. Tsonis, P. A. *et al.* A novel role of the hedgehog pathway in lens regeneration. *Dev. Biol.* **267**, 450–461 (2004).
51. Imokawa, Y. & Yoshizato, K. Expression of Sonic hedgehog gene in regenerating newt limb blastemas recapitulates that in developing limb buds. *Proc. Natl Acad. Sci. USA* **94**, 9159–9164 (1997).
52. Roy, S. & Gardiner, D. M. Cyclopamine induces digit loss in regenerating axolotl limbs. *J. Exp. Zool.* **293**, 186–190 (2002).
53. Laforest, L. *et al.* Involvement of the sonic hedgehog, patched 1 and bmp2 genes in patterning of the zebrafish dermal fin rays. *Development* **125**, 4175–4184 (1998).
54. Quint, E. *et al.* Bone patterning is altered in the regenerating zebrafish caudal fin after ectopic expression of sonic hedgehog and bmp2b or exposure to cyclopamine. *Proc. Natl Acad. Sci. USA* **99**, 8713–8718 (2002).
55. Liu, B. Y., McDermott, S. P., Khwaja, S. S. & Alexander, C. M. The transforming activity of Wnt effectors correlates with their ability to induce the accumulation of mammary progenitor cells. *Proc. Natl Acad. Sci. USA* **101**, 4158–4163 (2004).
56. Shackel, N. A., McGuinness, P. H., Abbott, C. A., Gorrell, M. D. & McCaughan, G. W. Identification of novel molecules and pathogenic pathways in primary biliary cirrhosis: cDNA array analysis of intrahepatic differential gene expression. *Gut* **49**, 565–576 (2001).
57. Monga, S. P., Padiaditakis, P., Mule, K., Stolz, D. B. & Michalopoulos, G. K. Changes in WNT/beta-catenin pathway during regulated growth in rat liver regeneration. *Hepatology* **33**, 1098–1109 (2001).
58. Surendran, K. & Simon, T. C. CNP gene expression is activated by Wnt signaling and correlates with Wnt4 expression during renal injury. *Am. J. Physiol. Renal Physiol.* **284**, F653–F662 (2003).
59. Miyaji, T. *et al.* Expression and distribution of transcripts for sonic hedgehog in the early phase of fracture repair. *Histochem. Cell Biol.* **119**, 233–237 (2003).
60. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
61. Dvorak, H. F. Tumors: wounds that do not heal. Similarities between tumour stroma generation and wound healing. *N. Engl. J. Med.* **315**, 1650–1659 (1986).
62. Wu, A. H., Wan, P. & Bernstein, L. A multiethnic population-based study of smoking, alcohol and body size and risk of adenocarcinomas of the stomach and esophagus (United States). *Cancer Causes Control* **12**, 721–732 (2001).
63. Lieber, C. S., Seitz, H. K., Garro, A. J. & Worner, T. M. Alcohol-related diseases and carcinogenesis. *Cancer Res.* **39**, 2863–2886 (1979).
64. Blair, A. & Kazerouni, N. Reactive chemicals and cancer. *Cancer Causes Control* **8**, 473–490 (1997).
65. Uemura, N. *et al.* *Helicobacter pylori* infection and the development of gastric cancer. *N. Engl. J. Med.* **345**, 784–789 (2001).
66. Ekblom, A., Helmick, C., Zack, M. & Adami, H. O. *Ulcerative colitis* and colorectal cancer. A population-based study. *N. Engl. J. Med.* **323**, 1228–1233 (1990).
67. Burak, K. *et al.* Incidence and risk factors for cholangiocarcinoma in primary sclerosing cholangitis. *Am. J. Gastroenterol.* **99**, 523–526 (2004).
68. Lessard, J. & Sauvageau, G. Bmi-1 determines the proliferative capacity of normal and leukaemic stem cells. *Nature* **423**, 255–260 (2003).
69. Park, I. K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423**, 302–305 (2003).
70. Molofsky, A. V. *et al.* Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation. *Nature* **425**, 962–967 (2003).
71. Leung, C. *et al.* Bmi1 is essential for cerebellar development and is overexpressed in human medulloblastomas. *Nature* **428**, 337–341 (2004).
72. Ito, H. *et al.* Hedgehog signaling molecules in bone marrow cells at the initial stage of fracture repair. *Biochem. Biophys. Res. Commun.* **262**, 443–451 (1999).
73. Hugh, T. J. *et al.* beta-catenin expression in primary and metastatic colorectal carcinoma. *Int. J. Cancer* **82**, 504–511 (1999).
74. Cano, A. *et al.* The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nature Cell Biol.* **2**, 76–83 (2000).
75. Yang, J. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumour metastasis. *Cell* **117**, 927–939 (2004).
76. Conacci-Sorrell, M. *et al.* Autoregulation of E-cadherin expression by cadherin-cadherin interactions: the roles of beta-catenin signaling, Slug, and MAPK. *J. Cell Biol.* **163**, 847–857 (2003).
77. Brabletz, T. *et al.* Variable beta-catenin expression in colorectal cancers indicates tumour progression driven by the tumour environment. *Proc. Natl Acad. Sci. USA* **98**, 10356–10361 (2001).
78. Wilson, A. J. & Gibson, P. R. Epithelial migration in the colon: filling in the gaps. *Clin. Sci. (Lond)* **93**, 97–108 (1997).
79. Kai, T. & Spradling, A. An empty *Drosophila* stem cell niche reactivates the proliferation of ectopic cells. *Proc. Natl Acad. Sci. USA* **100**, 4633–4638 (2003).
80. Watkins, D. N., Berman, D. M. & Baylin, S. B. Hedgehog signaling: progenitor phenotype in small-cell lung cancer. *Cell Cycle* **2**, 196–198 (2003).
81. Romer, J. T. *et al.* Suppression of the Shh pathway using a small molecule inhibitor eliminates medulloblastoma in *Ptc1*^{+/-} *p53*^{-/-} mice. *Cancer Cell* **6**, 229–240 (2004).
82. Tas, S. & Avci, O. Induction of the differentiation and apoptosis of tumour cells *in vivo* with efficiency and selectivity. *Eur. J. Dermatol.* **14**, 96–102 (2004).
83. Williams, J. A. *et al.* Identification of a small molecule inhibitor of the hedgehog signaling pathway: effects on basal cell carcinoma-like lesions. *Proc. Natl Acad. Sci. USA* **100**, 4616–4621 (2003).
84. Austin, T. W., Solar, G. P., Ziegler, F. C., Liem, L. & Matthews, W. A role for the *Wnt* gene family in hematopoiesis: expansion of multilineage progenitor cells. *Blood* **89**, 3624–3635 (1997).
85. van den Berg, D. J., Sharma, A. K., Bruno, E. & Hoffman, R. Role of members of the *Wnt* gene family in human hematopoiesis. *Blood* **92**, 3189–3202 (1998).
86. Taylor, M. D. *et al.* Mutations in SUFU predispose to medulloblastoma. *Nature Genet.* **31**, 306–311 (2002).
87. Kinzler, K. W. *et al.* Identification of an amplified, highly expressed gene in a human glioma. *Science* **236**, 70–73 (1987).
88. Dahmane, N. *et al.* The Sonic Hedgehog-Gli pathway regulates dorsal brain growth and tumorigenesis. *Development* **128**, 5201–5212 (2001).
89. Fan, L. *et al.* Hedgehog signaling promotes prostate xenograft tumour growth. *Endocrinology* **145**, 3961–3970 (2004).
90. Sanchez, P. *et al.* Inhibition of prostate cancer proliferation by interference with SONIC HEDGEHOG-GLI1 signaling. *Proc. Natl Acad. Sci. USA* **101**, 12561–12566 (2004).
91. Hamed, S. *et al.* Accelerated induction of bladder cancer in patched heterozygous mutant mice. *Cancer Res.* **64**, 1938–1942 (2004).
92. Nishimaki, H. *et al.* A role of activated Sonic hedgehog signaling for the cellular proliferation of oral squamous cell carcinoma cell line. *Biochem. Biophys. Res. Commun.* **314**, 313–320 (2004).
93. Derksen, P. W. *et al.* Illegitimate WNT signaling promotes proliferation of multiple myeloma cells. *Proc. Natl Acad. Sci. USA* **101**, 6122–6127 (2004).
94. Hoang, B. H. *et al.* Dickkopf 3 inhibits invasion and motility of Saos-2 osteosarcoma cells by modulating the Wnt-beta-catenin pathway. *Cancer Res.* **64**, 2734–2739 (2004).
95. Hoang, B. H. *et al.* Expression of LDL receptor-related protein 5 (LRP5) as a novel marker for disease progression in high-grade osteosarcoma. *Int. J. Cancer* **109**, 106–111 (2004).
96. You, L. *et al.* Inhibition of Wnt-2-mediated signaling induces programmed cell death in non-small-cell lung cancer cells. *Oncogene* **23**, 6170–6174 (2004).
97. Lee, A. Y. *et al.* Expression of the secreted frizzled-related protein gene family is downregulated in human mesothelioma. *Oncogene* **23**, 6672–6676 (2004).
98. Pola, R. *et al.* Postnatal recapitulation of embryonic hedgehog pathway in response to skeletal muscle ischemia. *Circulation* **108**, 479–485 (2003).
99. Moshiri, A. & Reh, T. A. Persistent progenitors at the retinal margin of *ptc*^{+/-} mice. *J. Neurosci.* **24**, 229–237 (2004).
100. Akazawa, C. *et al.* The upregulated expression of sonic hedgehog in motor neurons after rat facial nerve axotomy. *J. Neurosci.* **24**, 7923–7930 (2004).

Acknowledgements We thank E. Matunis for comments on the manuscript. Work in our laboratories is supported by the Howard Hughes Medical Institute, NIH, the Prostate Cancer Foundation and the Flight Attendant's Medical Research Institute. We apologise to authors of original work who could only be indirectly cited because of editorial constraints.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Stromal fibroblasts in cancer initiation and progression

Neil A. Bhowmick^{1,2,3}, Eric G. Neilson^{3,4} & Harold L. Moses^{1,3,4}

¹Department of Cancer Biology, ²Department of Urologic Surgery, ³The Vanderbilt-Ingram Cancer Center and ⁴Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA (e-mail: hal.moses@vanderbilt.edu)

It is widely accepted that the development of carcinoma — the most common form of human cancer — is due to the accumulation of somatic mutations in epithelial cells. The behaviour of carcinomas is also influenced by the tumour microenvironment, which includes extracellular matrix, blood vasculature, inflammatory cells and fibroblasts. Recent studies reveal that fibroblasts have a more profound influence on the development and progression of carcinomas than was previously appreciated. These new findings have important therapeutic implications.

The genetic lesions that occur in epithelial cells and lead to the initiation and progression of carcinomas have been defined during the past decade¹. Such discoveries of genetic changes in somatic cancer cells have not only advanced our basic understanding of tumour formation, but have also greatly influenced the treatment of cancer, with new therapies being targeted to specific pathways that are affected by genetic lesions (see introduction in this issue by Sawyers, page 294).

Carcinoma cells, like normal epithelial cells, live in a complex microenvironment that includes the extracellular matrix (ECM), diffusible growth factors and cytokines, and a variety of non-epithelial cell types, including those comprising the vasculature (endothelial cells, pericytes and smooth muscle cells), those that can respond to infection and injury (lymphocytes, macrophages and mast cells), and fibroblasts. It has long been recognized that carcinomas induce a modified stroma through the expression of growth factors that promote angiogenesis (the formation of new blood vessels), altered ECM expression, accelerated fibroblast proliferation, and increased inflammatory cell recruitment^{2,3} (Fig. 1).

Blood vessels are critical to the tumour microenvironment — without formation of new blood vessels, carcinomas cannot grow beyond a very small size, nor can they metastasize to colonize distant organs⁴. Tumour angiogenesis is due in part to secretion of endothelial growth factors by tumours, and indeed a targeted therapy (Avastin) that blocks the action of one of these factors (vascular endothelial growth factor or VEGF) has recently been approved⁵ (see introduction in this issue by Sawyers, page 294).

There is also a functional relationship between inflammation and cancer⁶. Cancers frequently arise in areas of chronic inflammation (see review in this issue by Beachy *et al.*, page 324). Examples include colon carcinoma, which is associated with inflammatory bowel disease, stomach cancer following *Helicobacter pylori* infection, and hepatocellular carcinomas after hepatitis C infection. Inflammatory cells, which arise independently of chronic inflammation, are also key to the microenvironment of carcinomas. Inflammatory cells probably influence cancer initiation and promotion by secreting cytokines, growth factors and chemokines, which stimulate proliferation of epithelia, and generate reactive oxygen species that can cause DNA damage⁶.

A crucial three-dimensional structure supports epithelia through the ECM, and impaired interactions of epithelial cells with ECM can result in the transformation of the epithelia into carcinoma^{7,8}. The specialized ECM that separates the epithelial and endothelial cells from the

stromal support components is termed the basement membrane. Whereas stromal ECM proteins are produced by fibroblasts, the major structural proteins of the basement membrane, including collagen IV, laminin, entactin and heparan-sulphate proteoglycans, are expressed by basal epithelia, myoepithelia and fibroblasts in a tissue-specific manner⁹. In turn, the unique composition of the basement membrane is thought to confer tissue specificity, epithelial polarity and functionality⁹.

Fibroblasts also have a well-recognized role in the carcinogenic process. They are responsible for the synthesis, deposition and remodelling of much of the ECM in tumour stroma (Fig. 1), and they are recognized as a source of paracrine (cell to cell) growth factors that influence the growth of carcinoma cells. However, fibroblasts have mostly been assumed to have a more passive role in cancer, responding to signals from the carcinoma cells. New data promote stromal fibroblasts from mere 'enablers' of cancer to potential 'inducers' of certain carcinomas. Here, we use recently reported examples to illustrate the influence of the stromal fibroblasts in epithelial neoplasia.

Overview of stromal-epithelial interactions

The importance of stromal interactions with epithelial cells is well established in embryonic development and tumorigenesis. The concept of a link between stromal cell maturation and adjacent epithelial proliferation was introduced more than 20 years ago¹⁰, and has gained support since^{11–16}. This interaction is mediated by soluble paracrine signals and ECM components secreted from developing mesenchyme that induce the adjacent epithelia to proliferate rapidly. As epithelial cells differentiate, so do adjacent stromal cells. These differentiated stromal cells generally express lower quantities of growth factors, and differentiated epithelia express cytokines for the maintenance of stromal differentiation, suggesting that a new balance of mesenchymal-epithelial crosstalk is reached during tissue maturation. During tumorigenesis, however, the prevailing model suggests a process whereby pre-cancerous epithelial cells acquire multiple genetic mutations¹⁷, and the associated stroma becomes 'activated', commonly expressing myofibroblastic markers^{2,3}. The characteristics of an activated carcinoma-associated fibroblast are not completely understood. However in our interpretation, such cells express α -smooth muscle actin, ECM proteins, and growth factors that act in an autocrine (within-cell) and paracrine fashion to potentiate and support the survival of a tumour.

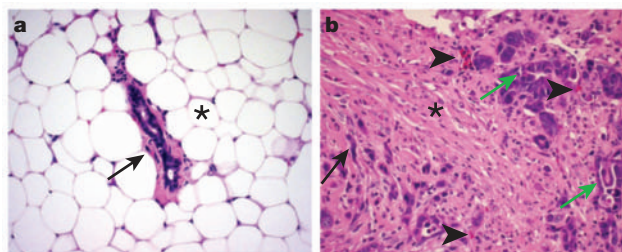


Figure 1 The stroma associated with the normal mammary gland differs profoundly from the stroma associated with a mammary carcinoma. **a**, The normal mammary gland has sparse connective tissue (arrow) surrounding the duct and abundant adipose tissue (asterisk). **b**, The carcinoma contains abundant connective tissue, probably as a result of growth-factor production by the carcinogenic environment. Note the dense collagen bundles associated with fibroblasts (asterisk) and the numerous small blood vessels and capillaries (arrow heads). The carcinoma cells form aberrant gland structures (green arrows), or grow in cords without gland formation (black arrow).

Emerging role of fibroblasts in epithelial cancer

Early evidence indicating an important role for stromal cells in cancer comes from tissue-culture experiments: epithelium from sub-mandibular glands transformed by polyoma virus does not grow when cultured by itself—it grows only in the presence of embryonic salivary gland mesenchyme¹⁸. The differences in the structure of normal stroma and tumour stroma are well known^{2,3} (Fig. 1). Apart from histological differences, tumour fibroblasts show enhanced proliferation and migratory behaviour *in vitro*^{19,20}. The constituents of the extracellular matrix and the vascular architecture in tumour stroma also differ from that associated with normal epithelia⁴.

These stromal events are mediated in part through autocrine growth-factor signalling involving the factors discussed above²¹. *In vitro* co-culture and *in vivo* xenograft systems demonstrate that factors derived from tumour fibroblasts contribute to the transformation of immortalized epithelia^{22,23}. In these studies, human fibroblasts from normal or tumour prostate were grown together with epithelial cells derived from benign prostate hyperplasia immortalized by the simian virus 40 (SV40) T-antigen. Tumour fibroblasts, but not

normal prostatic fibroblasts, stimulated epithelial proliferation and malignant transformation. However, tumour fibroblasts do not stimulate the growth of normal epithelial cells under identical conditions, suggesting that cancer-associated fibroblasts express ECM proteins and growth factors that influence the incipient tumour cells and promote the angiogenesis necessary to maintain epithelial transformation.

Experiments using irradiation of fibroblasts to cause sub-lethal DNA damage provide further evidence for a role for these cells in epithelial cancers. In mice, when non-transformed mammary epithelial cells were transplanted into fat pads containing irradiated fibroblasts, the incidence of breast tumours increased compared with transplant of similar epithelial cells into fat pads with fibroblasts that had not been irradiated²⁴. Recently, tissue recombination of pancreatic cancer cells with irradiated pancreatic fibroblasts resulted in a more aggressive and invasive cancer than when normal fibroblasts were used²⁵. Together, these reports demonstrate that changes in fibroblasts can contribute to epithelial transformation and more invasive behaviour. They also suggest that radiation-induced cancer may not only result from deleterious mutations in the epithelia, but also from alterations in stromal fibroblasts.

Interestingly, senescent human fibroblasts can promote the proliferation of pre-malignant and malignant epithelial cells in culture, and the formation of tumours in mice²⁶. This is probably due to the fact that senescent fibroblasts and cancer-associated fibroblasts express similar sets of paracrine growth factors that can contribute to cancer proliferation. Consistent with this theory, Krtolica *et al.*²⁶ found that even when senescent fibroblasts were only 10% of the fibroblast population, the epithelial cells proliferated. However, unlike the ability of cancer-associated fibroblasts to induce epithelial oncogenesis²², senescent fibroblasts are not able to convert non-transformed epithelia into carcinomas²⁶.

In a recent study of gene-expression profiles from each of the cell types present in normal breast tissue and breast carcinomas, changes in gene expression were observed in all cell types²⁷. Of particular interest was the overexpression of the chemokines CXCL14 and CXCL12 by myoepithelial cells and myofibroblasts, respectively. These chemokines were then shown to bind to receptors on the epithelial cells, causing enhanced proliferation and invasion. This study provides an elegant example of paracrine effects of factors derived from stromal cells on carcinoma cells.

Fibroblast-derived growth factors

Several families of growth factors, implicated as autocrine and paracrine mediators of stromal–epithelial interactions, are involved in carcinoma initiation and progression. These include the fibroblast growth factor (FGF) family, the insulin-like growth factor (IGF) family, the epithelial growth factor (EGF) family, hepatocyte growth factor (HGF), and the transforming growth factor- β (TGF- β) family (Table 1). Most of these factors are predominantly stimulators of proliferation and can play a part in promoting the carcinogenic process. The TGF- β family is different. With initial demonstrations that TGF- β could act as a growth inhibitor of most epithelial cells^{28,29}, it was speculated that this growth factor has a role in tumour suppression. Subsequent studies have indeed supported this hypothesis. For example, transgenic mouse studies demonstrated that increased TGF- β signalling can suppress tumour formation, whereas loss or attenuation of the pathway enhances carcinogenesis^{30–33}. In addition, inactivating mutations have been found in genes encoding components of the TGF- β signalling pathway in human tumours (see ref. 34 for a recent review). However, a loss of TGF- β sensitivity in carcinoma cells is frequently accompanied by increased expression of TGF- β by the same cells. Other sources of TGF- β in the microenvironment include the inflammatory cells and stromal fibroblasts³⁵. Furthermore, elevated levels of plasma TGF- β 1 can be detected in patients with cancer, and predicts early metastasis^{36–38}. The presence of TGF- β in the tumour

Table 1 Regulation of epithelial growth, differentiation and apoptosis

Soluble factors	Cells expressed	Responding cells	Possible role
HGF and MSP	Fibroblasts	Epithelia	+ Proliferation + Transformation + Morphogenic
IGF-1, IGF-2	Fibroblast	Epithelia (breast)	– Apoptosis + Proliferation
EGF and TGF- α	Epithelia and fibroblasts	Epithelia	+ Proliferation + Morphogenic
TGF- β 1, TGF- β 2, TGF- β 3	Epithelia and fibroblasts	Epithelia and fibroblasts	– Proliferation +/- Apoptosis + Morphogenic
FGF7/KGF	Fibroblast	Epithelia	+ Proliferation + Morphogenic
IL6, LIF, and oncostatin M	Fibroblast	Epithelia (colonic)	+ Proliferation + Transformation
FGF2	Fibroblast	Epithelia	+ Proliferation + Transformation
FGF10	Fibroblast	Epithelia	+ Proliferation
NGF	Fibroblast	Epithelia	+ Transformation
Stromal cell-derived factor 1 α (CXCL12)	Fibroblast	Epithelia (glioblastoma)	+ Proliferation + Transformation
Wnt1, Wnt3	Fibroblast	Epithelia	+ Proliferation + Transformation
MMP-1, MMP-7	Fibroblast	ECM and growth-factor activation in the stroma affect epithelia	+/- Proliferation +/- Apoptosis + Morphogenic

IL6, interleukin 6; LIF, leukaemia inhibitory factor; NGF, nerve growth factor.

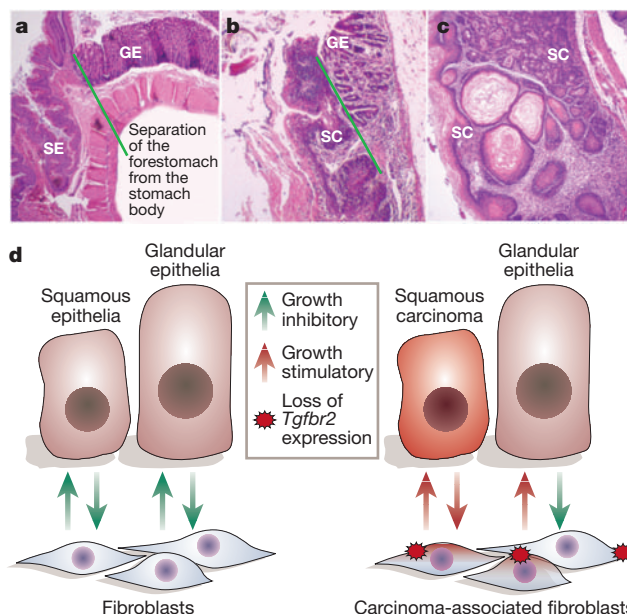
Box 1

Not all fibroblasts are created equal

A curious aspect of the *Tgfr2^{fspKO}* model³⁹ is that, apart from the prostate and the forestomach, other tissues in the mouse showed no signs of carcinogenic transformation. Functional differences between fibroblasts from different organs may explain this.

In a reductionist approach to understanding epithelial–stromal interactions, both in normal tissues and in cancer, the fibroblast component is often generalized and differences in fibroblasts from different tissues or differences within the same tissue are obscured. However, it is becoming clear that different fibroblasts can have distinct functions. For example, during lung development, epithelial induction capacities are known to differ depending on whether the mesenchyme is derived from the trachea or the tips of the growing lungs⁶⁶. The origin of the mesenchymal fibroblasts apparently determines their sensitivity to sonic hedgehog (Shh) derived from the developing lung epithelia. Shh signals through the patched receptor homologue (Ptch) found in the adjacent mesenchyme to stimulate proliferation⁶⁷. In turn, the specific mesenchyme supports the normal development and branching morphogenesis of the lung epithelia.

In mice, the fibroblasts associated with the squamous epithelium of the forestomach and oesophagus appear phenotypically similar to the fibroblasts associated with the adjacent glandular epithelia of the stomach body. Although a similar proportion of fibroblasts in the oesophagus, forestomach, and glandular stomach compartments were deficient for TGF- β signalling in *Tgfr2^{fspKO}* mice, only the squamous epithelia of the forestomach responded to the oncogenic paracrine signals. It is possible that the loss of TGF- β signalling in stromal fibroblasts results in proliferative autocrine and paracrine signals to which the fibroblasts and squamous epithelia of the forestomach uniquely responded with the formation of invasive carcinoma. Other epithelia may show no such response, for example, oesophagus epithelia showed no hyperplastic or neoplastic response in the *Tgfr2^{fspKO}* mice. Thus in this model, the paracrine signals required for malignancy of the glandular epithelia probably differ from that of the adjacent squamous epithelia.



Box 1 figure Stromal–epithelial interactions in normal and tumour tissues.

a, Wild-type stomach fibroblasts interacting with both the squamous epithelia (SE) of the forestomach and columnar glandular epithelia (GE) of the stomach body (green line indicates the area of epithelial transition). **b** and **c**, the progression of squamous cell carcinoma (SC) in the forestomach as it infiltrates the GE. The progression of the carcinoma is probably maintained through reciprocal signals to and from the fibroblasts. Fibroblasts have the capacity to proliferate and/or take on an activated form that can support and even initiate epithelial hyperplasia and eventual tumorigenesis. In turn, carcinoma-derived paracrine proliferative factors signal to the stroma. **d**, The non-carcinogenic balance is maintained by epithelia and fibroblasts as a result of dynamic interactions between the two compartments.

microenvironment may promote tumour growth by enhancing stromal support and angiogenesis and by impairing immune surveillance. In addition, while mutations in the TGF- β signalling pathway downstream of receptors may impair TGF- β -mediated growth inhibition, other pathway components may be retained, permitting TGF- β -mediated loss of adherens junctions, and increased motility of cancer cells — changes that would favour invasion and metastasis³⁴. Thus, TGF- β can exert both tumour suppressive and tumour promoting functions.

Genetic modification of fibroblasts can induce carcinoma

Recent studies provide evidence for a major role for TGF- β signalling in fibroblasts in the initiation of carcinomas. In one report, mice were generated in which the the TGF- β type II receptor gene (*Tgfr2*) was inactivated specifically in stromal fibroblasts using Cre-lox technology (resulting in *Tgfr2^{fspKO}* mice)³⁹. Here, exon 2 of *Tgfr2* was deleted in cells expressing FSP1 (fibroblast specific protein, S100A4). 100% of the mice showed prostatic intraepithelial neoplasia, a presumed forerunner of prostatic carcinoma, as well as invasive squamous cell carcinomas of the forestomach by six weeks of age³⁹. *Tgfr2^{fspKO}* fibroblasts overexpress HGF, and increases in the activating phosphorylation of the cognate HGF receptor, c-Met, were found in forestomach carcinoma cells. This suggests that activation of paracrine HGF signalling is one possible mechanism for the stimulation of epithelial proliferation. The *Tgfr2^{fspKO}* mouse model demonstrates that the TGF- β signalling pathway known to

suppress cell-cycle progression and tumour formation when acting on epithelial cells can also indirectly inhibit epithelial proliferation when acting in adjacent stromal fibroblasts *in vivo*. Consequently, loss of this pathway in fibroblasts results in increased epithelial proliferation and may also promote invasive carcinoma in some tissues.

More recently, mammary fibroblasts engineered to ectopically express HGF or TGF- β 1 alone or together induced mammary epithelia to develop ductal carcinoma *in situ*, adenocarcinoma, and poorly differentiated cancer, whereas transplantation of the same epithelial cell population with wild-type fibroblasts did not⁴⁰. Unique to this study was the transplant of human (rather than mouse) mammary epithelia and fibroblasts into cleared mammary fat pads of immune-deficient mice⁴⁰. This provides another example where modification of stromal fibroblasts can influence the malignant behaviour of adjacent epithelia.

HGF and TGF- β in epithelial and stromal cross talk

The stromally derived paracrine factors, HGF and TGF- β , have enjoyed the limelight in recent literature on epithelial–stromal crosstalk. Most cell types have the capacity to both express and respond to TGF- β (ref. 35). In contrast, HGF is primarily expressed by fibroblasts, while the cognate receptor, c-Met, is primarily expressed by epithelia⁴¹. Several reports support the transforming ability of HGF (ref. 42). As discussed above, the role of TGF- β is more complex and involves both tumour suppressive and tumour promoting roles.

The above mentioned report on tissue recombination of irradiated pancreatic fibroblasts with pancreatic cancer cells describes an elevated metastatic potential of the developing tumours²⁵. This phenomenon was associated with increased c-Met activation in the carcinoma cells and TGF- β 1 expression by the irradiated pancreatic fibroblasts. The fact that the authors did not observe a concomitant increase in HGF secretion by the fibroblasts suggests that the expression of an HGF-related ligand may be involved, such as MSP (macrophage stimulating protein). Alternatively, the overexpression of c-Met — a common finding in many cancers — may be associated with ligand-independent activation⁴³, or with increased sensitivity to physiological levels of HGF (ref. 44).

In light of the recent findings described above, apparent contradictions in our current understanding of the role of TGF- β in epithelial–stromal interactions have emerged. How can both overexpression of TGF- β with enhanced signalling, and loss of TGF- β signalling through knockout of *Tgfb2*, result in enhanced tumorigenesis? One possible answer to this apparent contradiction is that the overexpression of TGF- β by fibroblasts in mammary- and pancreas-transplantation models of cancer formation^{24,25,40} can affect both epithelial cells and stromal fibroblasts, whereas the loss of TGF- β signalling in the *Tgfb2*^{spKO} model affects TGF- β signalling only in fibroblasts. The enhanced HGF signalling in two mammary-tumour models^{24,40}, on the other hand, is probably the result of epithelial responses to paracrine factors (because c-Met receptors are primarily found in epithelial cells⁴³). Nevertheless, it is noteworthy that conditional deletion of *Tgfb2* in the epithelia of the mammary gland and prostate have no detectable phenotypic alterations (N.A.B, H.L.M. and A. Chytil, unpublished observations).

The *Tgfb2*^{spKO} mouse model³⁹ poses other important questions. If TGF- β stimulates an activated stromal environment, how does the loss of TGF- β signalling in stromal fibroblasts still allow for an otherwise activated stromal phenotype? The fact that tumours in the *Tgfb2*^{spKO} mice are associated with an activated stroma suggests important roles for factors other than TGF- β . As already discussed, a dual role for TGF- β as a tumour suppressor and promoter when acting on epithelial cells has emerged. However, in light of the above data^{39,40}, the paradigm of the early and late roles of TGF- β signalling on tumorigenesis needs to be adapted to include the influences of both fibroblasts and epithelia in tumour-susceptible tissues. In particular, both the direct effects of TGF- β on epithelia (through TGF- β receptors) and secondary effects (through the regulation of other growth factors) need to be considered.

Although the role of TGF- β in epithelia from various tissues has long been accepted as growth inhibitory and morphogenic through multiple downstream signalling proteins⁴⁵, the proliferative role of TGF- β in fibroblasts has been less clear as a result of the heterogeneity of fibroblasts (Box 1). Notably, NIH3T3 fibroblasts and dermal fibroblasts in culture are growth stimulated when treated with TGF- β (ref. 46). However, in the *Tgfb2*^{spKO} mice, the loss of TGF- β signalling in fibroblasts of the entire mouse³⁹ had little effect on fibroblast abundance in most tissues examined (that is, there was no evidence of stromal hyperplasia in the skin, lung, kidney, mammary gland, oesophagus, liver, small intestine or colon). This might indicate a tissue-selective role for TGF- β signalling in maintaining fibroblast homeostasis, or the presence of redundant growth inhibitory signals from other factors that do not require the TGF- β type II receptor. There was, however, significant stromal hyperplasia in the prostate and forestomach of *Tgfb2*^{spKO} mice — the same organs that undergo epithelial transformation³⁹. It is therefore important to keep in mind that multiple changes in several growth-factor pathways as well as tissue-specific responses will ultimately determine the outcome of the complex epithelial–fibroblast interactions in tumours compared to those in the non-diseased state (Box 1). In addition, our current knowledge of the role of many growth factors is primarily derived from studies of epithelial cells in culture. Undoubtedly, the ability to conditionally knock out growth-factor signalling specifically in

fibroblasts or epithelial cells will advance our understanding of how these networks of paracrine and autocrine signalling affect epithelial proliferation and transformation *in vivo*.

Other paracrine factors

The mechanisms involved in the microenvironmental effects on carcinoma progression are an area of intense investigation (Table 1); important insights have emerged in addition to the roles of TGF- β and HGF. Both the extracellular matrix and the matrix-degrading enzyme family of matrix metalloproteinases (MMP) can promote epithelial transformation⁴⁷. MMP levels and activities are often elevated in tumours. The activities of MMPs include pro-angiogenic and metastatic actions. They can also generate growth-regulating signals through the activation of growth factors, such as IGF (mediated through the cleavage of IGF-binding proteins), FGFs (mediated through the cleavage of perlecan), TGF- β and TGF- α (refs 47, 48). MMP-2 and MMP-9 knockout mice have a reduced susceptibility for lung metastases following intravenous injection of carcinoma cells^{48,49}. In addition, MMP-1 and MMP-7 are of fibroblastic origin and can induce increased susceptibility to mammary cancer when overexpressed in transgenic mice⁴⁷.

Platelet derived growth factor (PDGF) expressed by immortalized skin keratinocytes induces the expression of FGF7 (also known as keratinocyte growth factor or KGF) by fibroblasts⁵⁰. FGF7 in turn has been shown to produce further epithelial proliferation and promote carcinogenesis. In the prostate, FGF7 and FGF10, produced by fibroblasts, stimulate the proliferation of adjacent epithelia^{51,52}. This is countered by the paracrine growth factor FGF9, which is expressed by prostate epithelia, and received by fibroblasts⁵³. This is a clear example of how alterations in paracrine growth-factor pathways accompany the carcinogenesis process; the precise role of these factors in this process is still being elucidated.

In a recent study, the expression of PDGF by melanoma cells was shown to increase pericyte (special vascular cells) recruitment and proliferation in a B16 tumour model⁵⁴. The associated growth of the tumour was, however, not attributed to the increase in vasculature, but rather to pericyte-derived factors acting on the melanoma.

Another paracrine factor, Wnt1 (a known mammary epithelial oncogene), when expressed by fibroblasts initiates a morphological transformation of neighbouring C57MG mammary epithelial cells in co-culture experiments but no transformation of the Wnt1-expressing fibroblasts themselves is observed⁵⁵. More recently, Derksen *et al.*⁵⁶ provided supporting data on the paracrine role of WNT family proteins enhancing the survival and growth of multiple myeloma cells in humans. The paracrine dynamics in all these experiments is modulated by the presence or absence of contravening influences that are unevenly distributed in tissues.

The role of mutations in stromal cells

One of the more provocative implications of the recent mouse models by Bhowmick *et al.*³⁹ and Kuperwasser *et al.*⁴⁰ is that epithelial cells without specifically engineered mutations can be induced to form carcinomas by association with genetically altered fibroblasts. In previous studies, the epithelial cells induced to be more carcinogenic by association with irradiated fibroblasts were already fully transformed into carcinomas²⁵, or had known p53 gene mutations²⁴. Of course, it is possible that in the studies involving recombination of human mammary epithelial cells with human fibroblasts modified to express HGF and/or TGF- β 1 in cleared mammary fat pads of immune deficient mice⁴⁰ pre-existing mutations existed in the human mammary epithelial cells that formed carcinomas. This uncertainty is supported by the observation that only selected epithelial preparations gave carcinomas when recombined with HGF- and/or TGF- β -expressing fibroblasts. Most probably, however, the mammary epithelial cells forming carcinomas acquired tumorigenic changes subsequent to their interaction with the HGF- and/or TGF- β 1-expressing fibroblasts.

Likewise, the pre-neoplastic and neoplastic lesions observed in the *Tgfr2^{spKO}* mouse model³⁹ suggest that pre-malignant and malignant epithelial tumours can develop owing to mutations in fibroblasts preceding any subsequent tumorigenic changes in the epithelial cells. One possible explanation for this is that the rapid proliferation induced in epithelial cells by HGF (and probably other epithelial cell growth factors) secreted by the (TGF- β) type II receptor (T β RII)-null fibroblasts leads to formation of genetic lesions (Fig. 2). Another possible explanation, which is not mutually exclusive with the first, is the generation of oxygen free radicals or other endogenous mutagens — a process that could be accelerated by the recruitment of inflammatory cells to involved sites⁵⁷. The selective presence and timing of putative genetic alterations in some epithelial cells in the *Tgfr2^{spKO}* model remains to be determined.

Earlier studies have demonstrated mutations in stromal fibroblasts that accompany carcinomas. Loss of heterozygosity (LOH) has

been found with high frequency in stromal cells associated with human breast cancer⁵⁸. Furthermore, mutations in two tumour suppressor genes, *p53* and *PTEN*, were demonstrated in both mammary carcinoma cells and associated stromal fibroblasts⁵⁹. However, the possibility that the stromal compartment contains fibroblasts derived from carcinoma cells that have undergone an epithelial to mesenchymal transition cannot be excluded⁶⁰.

Additional evidence for a role of mutations in stromal cells in the development of epithelial hyperplasias comes from studies of heritable juvenile polyposis syndrome, which is characterized by an overgrowth of tissue native to the area in which they normally occur. In this condition, the loss of normal genetic function occurs predominantly in interstitial fibroblasts among colonic polyps^{61,62}. It is of interest that the genes targeted in this syndrome encode bone morphogenetic protein (BMP) receptor 1A (BMPRI1A), a member of the TGF- β family of receptors, or the tumour suppressor SMAD4 which has a central role in signalling from both the BMP and TGF- β receptors⁶³.

These data, along with those reported by Bhowmick *et al.*³⁹, strongly support the hypothesis that TGF- β superfamily signalling in stromal fibroblasts frequently exerts a tumour suppressive function on adjacent epithelia. This is of particular interest because TGF- β -signalling pathways within epithelial cells are also tumour suppressive³⁴. These findings suggest that normal cells *in vivo* restrict the malignant phenotype of neighbouring cells, and that initiation and progression of carcinomas probably involves the overcoming of constraints from normal interstitial tissue through a combination of potential genetic, epigenetic and stromal changes.

Implications for therapy

Because of the positive role TGF- β signalling has in carcinomas, such as causing excess fibrosis, tumour progression and metastasis, the pharmaceutical industry has been developing inhibitors of TGF- β signalling pathways. Indeed, studies involving systemic inhibition of TGF- β signalling in adult animals has demonstrated no adverse effects and has decreased metastases from mammary carcinomas⁶⁴. However, the more recent studies described above raise the question: can such inhibitors promote carcinomas through the inhibition of TGF- β signalling in normal stroma? If this is the case, caution is warranted in pursuing this approach.

Another therapeutic target suggested by the data discussed in this review is HGF. As pointed out by Ohuchida *et al.*²⁵, irradiation of stromal cells can cause activation of c-Met in carcinoma cells. This group further reported that specific antagonists of HGF could block the enhanced invasiveness of pancreatic carcinoma caused by irradiated fibroblasts. Overexpression and activation of c-Met is a common event in human cancer⁶⁵, and a recent publication reported the development of a soluble c-Met receptor (decoy Met) that interferes with HGF binding to c-Met and c-Met homodimerization⁴³. Local and systemic delivery of decoy Met significantly inhibited proliferation and metastasis in human tumour xenografts.

Conclusions and future directions

The studies discussed in this review suggest a more important role for stromal fibroblasts in carcinogenesis than was previously appreciated. Fibroblasts influence epithelial transformation by producing paracrine factors that affect both normal epithelia as well as carcinoma cells. In addition, there is now considerable evidence that mutations arising in stromal fibroblasts can precede carcinoma development. The tissue specificity of stromal–epithelial interactions probably accounts for a tissue- and cell-type specific role of the microenvironment in carcinoma development. Importantly, some of the data point to potential targets for therapy, specifically inhibition of the HGF–c-Met axis. Here, we have focused on the effect of various stroma-derived paracrine factors on epithelial cells. It is likely that these same factors also have important effects on stromal cells, including fibroblasts, endothelial cells and inflammatory cells.

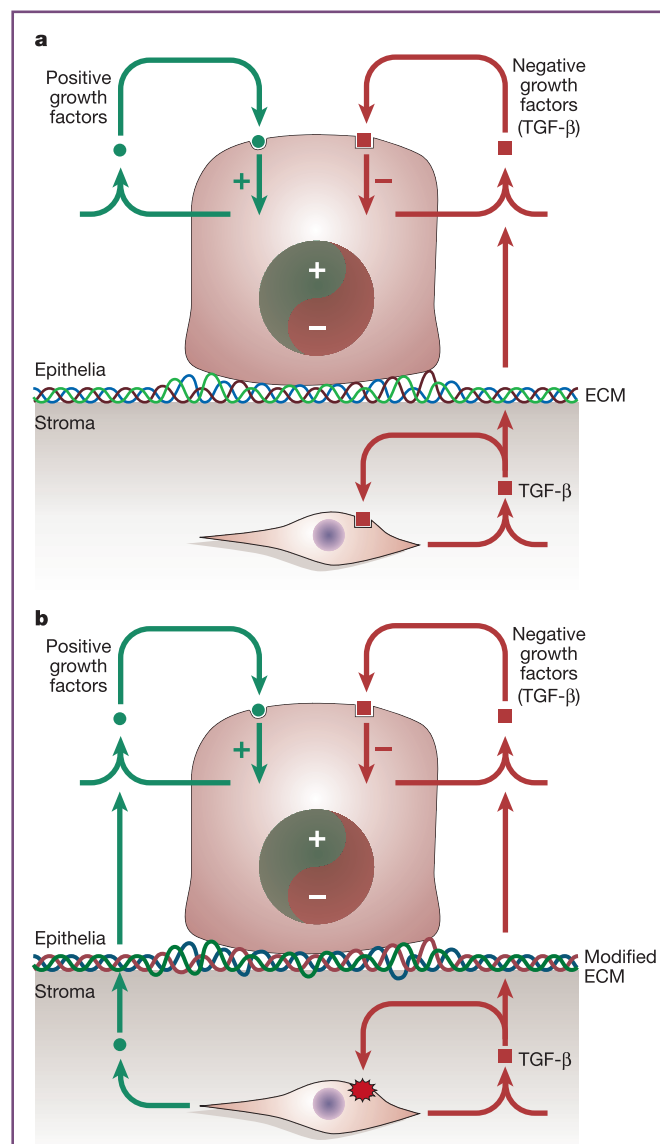


Figure 2 Epithelia can be reactive to a changing stromal environment. **a**, Homeostatic interactions between the epithelia and fibroblasts are maintained by positive and negative signals that influence the proliferation and differentiation of both the stroma and epithelia. **b**, When signalling by a suppressive growth factor (TGF- β) to the stromal fibroblasts is lost (red starburst), it leads to elevated fibroblast proliferation. Resulting paracrine factors (for example, HGF) and potential modifications in the ECM can stimulate the proliferation and transformation of epithelial cells *in vivo* in some tissues.

Valuable insights concerning the tissue microenvironment have been derived from co-culture and tissue recombination xenograph experiments, but such information may not be applicable to the *in vivo* situation because not all important environmental factors and cells are considered. The ability to overexpress specific factors or conditionally knock out specific genes *in vivo* in tissue fibroblasts and other cells of the normal stroma and tumour microenvironment will add greatly to our knowledge of the complex interactions involved in tissue homeostasis as well as those changes involved in the initiation and progression of cancer. □

doi:10.1038/nature03096

1. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
2. Ronnov-Jessen, L., Petersen, O. W. & Bissell, M. J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
3. Tlsty, T. D. & Hein, P. W. Know thy neighbor: stromal cells can contribute oncogenic signals. *Curr. Opin. Genet. Dev.* **11**, 54–59 (2001).
4. Carmeliet, P. & Jain, R. K. Angiogenesis in cancer and other diseases. *Nature* **407**, 249–257 (2000).
5. Sandler, A. B., Johnson, D. H. & Herbst, R. S. Anti-vascular endothelial growth factor monoclonals in non-small cell lung cancer. *Clin. Cancer Res.* **10**, 4258s–4262s (2004).
6. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
7. Weaver, V. M. *et al.* Reversion of the malignant phenotype of human breast cells in three-dimensional culture and *in vivo* by integrin blocking antibodies. *J. Cell Biol.* **137**, 231–245 (1997).
8. Ronnov-Jessen, L., Petersen, O. W., Kotliarsky, V. E. & Bissell, M. J. The origin of the myofibroblasts in breast cancer. Recapitulation of tumour environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. *J. Clin. Invest.* **95**, 859–873 (1995).
9. Kalluri, R. Basement membranes: structure, assembly and role in tumour angiogenesis. *Nature Rev. Cancer* **3**, 422–433 (2003).
10. Cunha, G. R., Reese, B. A. & Sekkingstad, M. Induction of nuclear androgen-binding sites in epithelium of the embryonic urinary bladder by mesenchyme of the urogenital sinus of embryonic mice. *Endocrinology* **107**, 1767–1770 (1980).
11. Xin, L., Ide, H., Kim, Y., Dube, P. & Witte, O. N. *In vivo* regeneration of murine prostate from dissociated cell populations of postnatal epithelia and urogenital sinus mesenchyme. *Proc. Natl Acad. Sci. USA* **100** (suppl. 1), 11896–11903 (2003).
12. Cunha, G. R., Bigsby, R. M., Cooke, P. S. & Sugimura, Y. Stromal-epithelial interactions in adult organs. *Cell Differ.* **17**, 137–148 (1985).
13. Culig, Z. *et al.* Regulation of prostatic growth and function by peptide growth factors. *Prostate* **28**, 392–405 (1996).
14. Cunha, G. R. & Donjacour, A. A. Mesenchymal-epithelial interactions in the growth and development of the prostate. *Cancer Treat. Res.* **46**, 159–75 (1989).
15. Hayward, S. W. Approaches to modeling stromal-epithelial interactions. *J. Urol.* **168**, 1165–1172 (2002).
16. Hayward, S. W. & Cunha, G. R. The prostate: development and physiology. *Radiol. Clin. North Am.* **38**, 1–14 (2000).
17. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nature Med.* **10**, 789–799 (2004).
18. Dawe, C. J. *In Tissue Interactions in Carcinogenesis*. (ed. Tarin, D.) 305–358 (Academic, London, 1972).
19. Schor, S. L., Schor, A. M. & Rushton, G. Fibroblasts from cancer patients display a mixture of both foetal and adult-like phenotypic characteristics. *J. Cell Sci.* **90**, 401–407 (1988).
20. Schor, S. L., Schor, A. M., Rushton, G. & Smith, L. Adult, foetal and transformed fibroblasts display different migratory phenotypes on collagen gels: evidence for an isoformic transition during foetal development. *J. Cell Sci.* **73**, 221–234 (1985).
21. Russell, P. J., Bennett, S. & Stricker, P. Growth factor involvement in progression of prostate cancer. *Clin. Chem.* **44**, 705–723 (1998).
22. Olumi, A. F. *et al.* Carcinoma-associated fibroblasts direct tumour progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–11 (1999).
23. Hayward, S. W. *et al.* Malignant transformation in a nontumorigenic human prostatic epithelial cell line. *Cancer Res.* **61**, 8135–8142 (2001).
24. Barcellos-Hoff, M. H. & Ravan, S. Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. *Cancer Res.* **60**, 1254–1260 (2000).
25. Ohuchida, K. *et al.* Radiation to stromal fibroblasts increases invasiveness of pancreatic cancer cells through tumour–stromal interactions. *Cancer Res.* **64**, 3215–3222 (2004).
26. Krtolica, A., Parrinello, S., Lockett, S., Desprez, P. Y. & Campisi, J. Senescent fibroblasts promote epithelial cell growth and tumorigenesis: a link between cancer and aging. *Proc. Natl Acad. Sci. USA* **98**, 12072–12077 (2001).
27. Allinen, M. *et al.* Molecular characterization of the tumour microenvironment in breast cancer. *Cancer Cell* **6**, 17–32 (2004).
28. Tucker, R. F., Shipley, G. D., Moses, H. L. & Holley, R. W. Growth inhibitor from BSC-1 cells closely related to platelet type beta transforming growth factor. *Science* **226**, 705–777 (1984).
29. Moses, H. L. *et al.* Type β transforming growth factor is a growth stimulator and a growth inhibitor. *Cancer Cells* **3**, 65–71 (1985).
30. Pierce, D. F. *et al.* Mammary tumour suppression by transforming growth factor beta 1 transgene expression. *Proc. Natl Acad. Sci. USA* **92**, 4254–4258 (1995).
31. Bottinger, E. P., Jakubczak, J. L., Haines, D. C., Bagnall, K. & Wakefield, L. M. Transgenic mice overexpressing a dominant-negative mutant type II transforming growth factor beta receptor show enhanced tumorigenesis in the mammary gland and lung in response to the carcinogen 7, 12-dimethylbenz[*a*]-anthracene. *Cancer Res.* **57**, 5564–5570 (1997).
32. Gorska, A. E. *et al.* Transgenic mice expressing a dominant-negative mutant type II TGF- β receptor have impaired mammary development and enhanced mammary tumour formation. *Am. J. Pathol.* **163**, 1539–1549 (2003).
33. Amendt, C., Schirmacher, P., Weber, H. & Blessing, M. Expression of a dominant negative type II

- TGF-beta receptor in mouse skin results in an increase in carcinoma incidence and an acceleration of carcinoma development. *Oncogene* **17**, 25–34 (1998).
34. Akhurst, R. J. & Derynck, R. TGF-beta signalling in cancer—a double-edged sword. *Trends Cell Biol.* **11**, S44–S51 (2001).
35. Gold, L. I. The role for transforming growth factor-beta (TGF-beta) in human cancer. *Crit. Rev. Oncog.* **10**, 303–360 (1999).
36. Tsumihama, H. *et al.* Circulating transforming growth factor beta 1 as a predictor of liver metastasis after resection in colorectal cancer. *Clin. Cancer Res.* **7**, 1258–1262 (2001).
37. Shariat, S. F. *et al.* Preoperative plasma levels of transforming growth factor beta(1) (TGF-beta(1)) strongly predict progression in patients undergoing radical prostatectomy. *J. Clin. Oncol.* **19**, 2856–2864 (2001).
38. Shariat, S. F. *et al.* Preoperative plasma levels of transforming growth factor beta(1) strongly predict clinical outcome in patients with bladder carcinoma. *Cancer* **92**, 2985–2992 (2001).
39. Bhowmick, N. A. *et al.* TGF-beta signalling in fibroblasts modulates the oncogenic potential of adjacent epithelia. *Science* **303**, 848–851 (2004).
40. Kuperwasser, C. *et al.* Reconstruction of functionally normal and malignant human breast tissues in mice. *Proc. Natl Acad. Sci. USA* **101**, 4966–4971 (2004).
41. Nakamura, T., Matsumoto, K., Kiritoshi, A. & Tano, Y. Induction of hepatocyte growth factor in fibroblasts by tumour-derived factors affects invasive growth of tumour cells: *in vitro* analysis of tumour–stromal interactions. *Cancer Res.* **57**, 3305–3313 (1997).
42. Birchmeier, C., Birchmeier, W., Gherardi, E. & Vande Woude, G. F. Met, metastasis, motility and more. *Nature Rev. Mol. Cell Biol.* **4**, 915–925 (2003).
43. Michieli, P. *et al.* Targeting the tumour and its microenvironment by a dual-function decoy Met receptor. *Cancer Cell* **6**, 61–73 (2004).
44. Pennacchietti, S. *et al.* Hypoxia promotes invasive growth by transcriptional activation of the met protooncogene. *Cancer Cell* **3**, 347–361 (2003).
45. Massague, J., Blain, S. W. & Lo, R. S. TGF-beta signalling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).
46. Koskinen, P. J., Sistonen, L., Bravo, R. & Alitalo, K. Immediate early gene responses of NIH 3T3 fibroblasts and NMuMG epithelial cells to TGF beta-1. *Growth Factors* **5**, 283–293 (1991).
47. Lynch, C. C. & Matrisian, L. M. Matrix metalloproteinases in tumour-host cell communication. *Differentiation* **70**, 561–573 (2002).
48. Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nature Rev. Cancer* **2**, 161–174 (2002).
49. Itoh, T. *et al.* Reduced angiogenesis and tumour progression in gelatinase A-deficient mice. *Cancer Res.* **58**, 1048–1051 (1998).
50. Brauchle, M., Angermeyer, K., Hubner, G. & Werner, S. Large induction of keratinocyte growth factor expression by serum growth factors and pro-inflammatory cytokines in cultured fibroblasts. *Oncogene* **9**, 3199–3204 (1994).
51. Yan, G., Fukabori, Y., Nikolopoulos, S., Wang, F. & McKeen, W. L. Heparin-binding keratinocyte growth factor is a candidate stromal-to-epithelial-cell andromedin. *Mol. Endocrinol.* **6**, 2123–2128 (1992).
52. Lu, W., Luo, Y., Kan, M. & McKeen, W. L. Fibroblast growth factor-10. A second candidate stromal to epithelial cell andromedin in prostate. *J. Biol. Chem.* **274**, 12827–12834 (1999).
53. Jin, C. *et al.* Directionally specific paracrine communication mediated by epithelial FGF9 to stromal FGFR3 in two-compartment premalignant prostate tumours. *Cancer Res.* **64**, 4555–4562 (2004).
54. Furuhashi, M. *et al.* Platelet-derived growth factor production by B16 melanoma cells leads to increased pericyte abundance in tumours and an associated increase in tumour growth rate. *Cancer Res.* **64**, 2725–2733 (2004).
55. Jue, S. F., Bradley, R. S., Rudnicki, J. A., Varmus, H. E. & Brown, A. M. The mouse Wnt-1 gene can act via a paracrine mechanism in transformation of mammary epithelial cells. *Mol. Cell. Biol.* **12**, 321–328 (1992).
56. Derksen, P. W. *et al.* Illegitimate WNT signalling promotes proliferation of multiple myeloma cells. *Proc. Natl Acad. Sci. USA* **101**, 6122–6127 (2004).
57. Marnett, L. J., Riggins, J. N. & West, J. D. Endogenous generation of reactive oxidants and electrophiles and their reactions with DNA and protein. *J. Clin. Invest.* **111**, 583–593 (2003).
58. Moir, F. *et al.* Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis. *Cancer Res.* **60**, 2562–2566 (2000).
59. Kurose, K. *et al.* Frequent somatic mutations in PTEN and TP53 are mutually exclusive in the stroma of breast carcinomas. *Nature Genet.* **32**, 355–357 (2002).
60. Xue, C., Plith, D., Venkov, C., Xu, C. & Neilson, E. G. The gatekeeper effect of epithelial–mesenchymal transition regulates the frequency of breast cancer metastasis. *Cancer Res.* **63**, 3386–3394 (2003).
61. Jacoby, R. F. *et al.* A juvenile polyposis tumour suppressor locus at 10q22 is deleted from non-epithelial cells in the lamina propria. *Gastroenterology* **112**, 1398–1403 (1997).
62. Wirtzfeld, D. A., Petrelli, N. J. & Rodriguez-Bigas, M. A. Hamartomatous polyposis syndromes: molecular genetics, neoplastic risk, and surveillance recommendations. *Ann. Surg. Oncol.* **8**, 319–327 (2001).
63. Waite, K. A. & Eng, C. From developmental disorder to heritable cancer: it's all in the BMP/TGF-beta family. *Nature Rev. Genet.* **4**, 763–773 (2003).
64. Yang, Y. A. *et al.* Lifetime exposure to a soluble TGF-beta antagonist protects mice against metastasis without adverse side effects. *J. Clin. Invest.* **109**, 1607–1615 (2002).
65. Trusolino, L. & Comoglio, P. M. Scatter-factor and semaphorin receptors: cell signalling for invasive growth. *Nature Rev. Cancer* **2**, 289–300 (2002).
66. Wessells, N. K. Mammalian lung development: interactions in formation and morphogenesis of tracheal buds. *J. Exp. Zool.* **175**, 455–466 (1970).
67. Bellucci, S. *et al.* Involvement of Sonic hedgehog (Shh) in mouse embryonic lung growth and morphogenesis. *Development* **124**, 53–63 (1997).

Acknowledgements This work was supported by a DOD USAMRMC grant (N.A.B.), NIH grants (H.L.M.), and a Vanderbilt-Ingram Cancer Center Support Grant. We thank J. Pietenpol and S. Hayward for critical review of the manuscript and helpful advice.

Competing interests statement The authors declare that they have no competing financial interests.

Aneuploidy and cancer

Harith Rajagopalan & Christoph Lengauer

The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, 1650 Orleans Street, Baltimore, Maryland 21231, USA (e-mail: lengauer@jhmi.edu)

In contrast to normal cells, aneuploidy — alterations in the number of chromosomes — is consistently observed in virtually all cancers. A growing body of evidence suggests that aneuploidy is often caused by a particular type of genetic instability, called chromosomal instability, which may reflect defects in mitotic segregation in cancer cells. A better understanding of the molecular mechanisms leading to aneuploidy holds promise for the development of cancer drugs that target this process.

A fundamental principle underlying our understanding of tumorigenesis is that cancers arise from the sequential acquisition of genetic alterations in specific genes^{1,2}. These changes (mutations and amplifications, for instance) occur in individual cells within a population: each change attains fixation from a wave of clonal expansion due to the relative growth advantage that the new mutation confers on the cell. In this way, genetic events represent rate-limiting bottlenecks in the clonal evolution of cancers. But the number of bottlenecks needed for a single cell to develop into a metastatic cancer is not clear. Mathematical extrapolations have suggested that most cancers require six to ten such clonal events to fully mature^{3,4}.

In this context, let us consider one of the most common properties of cancers — aneuploidy. More than a century ago, one of the first things that researchers noticed when they looked at cancer cells under the microscope was that tumour cells often had excess chromosomes^{5,6}. Normal human cells (even those invading and immediately surrounding cancers) contain an invariable complement of 46 chromosomes. However, most cancers contain cells that not only possess an abnormal number of chromosomes (often between 60 and 90) but that also differ from each other in the number of chromosomes they contain. Furthermore, these chromosomes commonly have structural aberrations that are vanishingly rare in normal cells: inversions, deletions, duplications, and translocations (Fig. 1). These numerical and structural abnormalities define aneuploidy.

Why are cancers aneuploid? Because it is so common, it has been suggested that aneuploidy, like defects in signalling pathways, is essential for tumorigenesis^{7,8}. Alternatively, aneuploidy may be a meaningless consequence of deregulated growth in cancer cells; simply a cytogenetic silhouette of other processes that are essential for tumorigenesis. In addition to the question of aneuploidy's relevance to tumorigenesis, its genetic basis is also unclear. Although many believe that aneuploidy can be caused by genetic changes within tumours, others have suggested that aneuploidy is entirely independent of gene mutations⁷. To tease apart these possibilities, we return to observations on specific cancers and attempt to draw implications from these studies. We review historical and modern efforts to explain the causes and consequences of aneuploidy and its contribution to cancer. Finally, we consider the therapeutic implications offered by these findings.

Defects in mitosis

Mitotic defects in tumour cells with aneuploidy were first described by David Hanseman more than a century ago⁹. In tissue sections from various carcinomas, he encountered mitotic figures (chromosome architectures) that were

abnormal in size and structure. In particular, he described asymmetrical mitoses in cells displaying bipolar anaphase and telophase. Here, the two groups of segregating sister chromatids contained unequal amounts of chromatin. Hanseman also described two additional chromosomal structures that often coincided with asymmetrical mitoses: the formation of anaphase bridges and multipolar mitoses. All three processes can result in abnormalities in chromosome numbers and to a gradual loss of heterozygosity by sequential loss of mitotically unstable chromosomes^{10,11}.

However, a significant proportion of solid tumours show polyploid chromosome numbers (duplication of a full chromosome complement), which would not be expected to arise from a gradual loss of individual chromosomes from a diploid cell. Studies of the cytogenetic evolution in breast cancer have suggested that a highly aneuploid state could originate from a tetraploidization (whole genome duplication) event concurrent with a gradual loss of individual chromosome copies¹². This idea is supported by the finding of a high frequency of tetraploid cells in certain pre-neoplastic lesions, such as Barrett's oesophagus¹³ and ulcerative colitis¹⁴. Further evidence stems from the observation that aneuploid tumours often show a duplication of some of their structurally rearranged chromosomes, including balanced translocations. The precise mechanisms behind such a whole-genome duplication remain to be elucidated. Recent experiments suggest that short telomeres could have a key role, as telomerase-negative immortalized cells tend to develop a tetraploid cell population¹⁵.

When a tumour cell acquires double the number of chromosomes, it also acquires double the number of centrosomes. This raises the question of whether abnormal centrosome numbers in tumours are merely side effects of tetraploidization or whether they have a causal role in tumour evolution. An *in vitro* model overexpressing the kinase Aurora A has shown that tetraploidization is a major route to centrosome amplification¹⁶. Approximately 80% of invasive breast tumours show abnormalities in the structure and/or number of centrosomes, and abnormalities in centrosome size and centrosome number show a positive correlation with aneuploidy¹⁷. Centrosomal abnormalities have also been detected at the breast carcinoma *in situ* stage, indicating that they develop early in the neoplastic process¹⁸.

Lessons from cancers

Moving away from a descriptive approach to understanding aneuploidy, more recent work has focused on understanding the teleological basis for why cancers develop this ubiquitous property. Ironically, new insights into aneuploidy have come from the analysis of the small fraction of cancers that are not aneuploid. Approximately 15% of colorectal cancers show a form of genetic instability that is

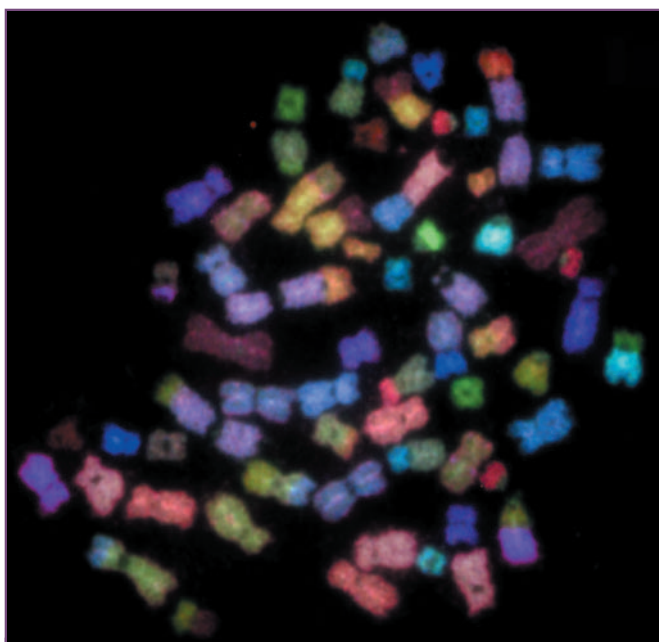


Figure 1 Multicolour-fluorescence *in situ* hybridization of an aneuploid non-small cell lung cancer. Metaphase chromosomes are stained in 23 different colours (provided by M. R. Speicher).

characterized by mismatch repair (MMR) deficiency. This is generally due to inactivation of the MMR genes *hMLH1* or *hMSH2* (ref. 19). Loss of MMR function renders tumour cells susceptible to the acquisition of somatic mutations throughout the genome. Simple repeat sequences are particularly susceptible to mutations in the absence of MMR, and the genetic instability in these tumours is often referred to as microsatellite instability (MIN or MSI)^{20–22}. Cancer cells that possess MIN have a mutation rate at the nucleotide level that is two to three orders of magnitude greater than that observed in normal cells. However, these cancer cells retain a diploid or near-diploid chromosome content²³.

Another example of cancers without aneuploidy is the hereditary cancer syndrome Xeroderma pigmentosum (XP). XP patients, who inherit a deficiency in one of the genes in the nucleotide excision repair (NER) pathway, develop skin cancers whose genomes are characterized by a high mutation rate at pyrimidine dimers. Unlike sporadic skin cancers (basal cell carcinomas, squamous cell carcinomas, and melanomas) that develop in non-XP patients, NER-deficient skin cancers are not aneuploid²⁴.

To summarize, although most solid tumours are aneuploid, the few that do not show aneuploidy seem to have inactivated specific DNA-repair pathways. The implication is that these defects allow cancers to accelerate their mutation rates. Indeed, an elevated mutation rate may be necessary to allow cancers to acquire the numerous genetic changes required for tumorigenesis. Given the mutual exclusivity of DNA-repair inactivation and aneuploidy, the corollary argument is that aneuploidy may reflect a different form of genetic instability which accelerates the accumulation of mutations, albeit by different mechanisms.

Aneuploidy and chromosomal instability

One attempt to understand the processes responsible for aneuploidy involved measuring the rate at which colon cancer cell lines gain and lose chromosomes²⁵. Clones were generated and expanded through a defined number of generations before they were examined by fluorescence *in situ* hybridization with DNA probes specific for centromeric regions of individual chromosomes. In cell lines that did not show MIN, the probability of losing or gaining a chromosome was 0.01 per chromosome per cell division. The corresponding rate in MIN-cell

lines was much lower and could not be accurately determined. This accelerated rate of chromosomal gains and losses that can lead to aneuploidy was termed chromosomal instability (CIN), and has been suggested as an alternative form of genetic instability. What benefit might chromosomal instability confer on a dividing cancer cell? Conceivably, CIN could accelerate the rate of loss of heterozygosity of a tumour suppressor gene and/or effectively amplify an oncogene by duplicating the chromosome on which it lies, thus contributing to tumorigenesis.

Many cancer-cell chromosomes also possess structural abnormalities, including interstitial deletions, inversions and translocations, but it remains unclear whether these changes occur at a higher rate in cancer cells than in normal cells. These structural changes are not measured by the CIN assay described above. It is useful, however, in that it suggests that defective chromosomal segregation leading to CIN may be key to understanding aneuploidy. It should also be noted that the invocation of chromosomal instability as the cause of aneuploidy in cancer is simply one hypothesis.

Molecular mechanisms of chromosomal instability

Assuming that chromosomal instability is one path leading to aneuploidy, the next challenge is to try to understand the mechanisms underlying chromosomal instability. Mounting evidence implicates the mitotic spindle checkpoint as the point of failure in CIN (ref. 25). The normal function of the spindle checkpoint is to ensure that all chromosomes are correctly aligned in metaphase cells and properly attached to the mitotic spindle before chromosome separation can proceed. Chromosomally unstable cells in tissue culture do not arrest in mitosis as well as normal cells when they are incubated with microtubule-disrupting agents. A partially compromised checkpoint may represent a balance that weighs an improved fitness in changing environments against the possibility of apoptosis from numerous genetic insults²⁶.

Because of the role the spindle checkpoint seems to have in CIN, mutations in putative CIN genes may be identified by examining the components of the kinetochore and spindle checkpoint apparatus (Fig. 2). The ever-expanding list of proteins that have a role in these

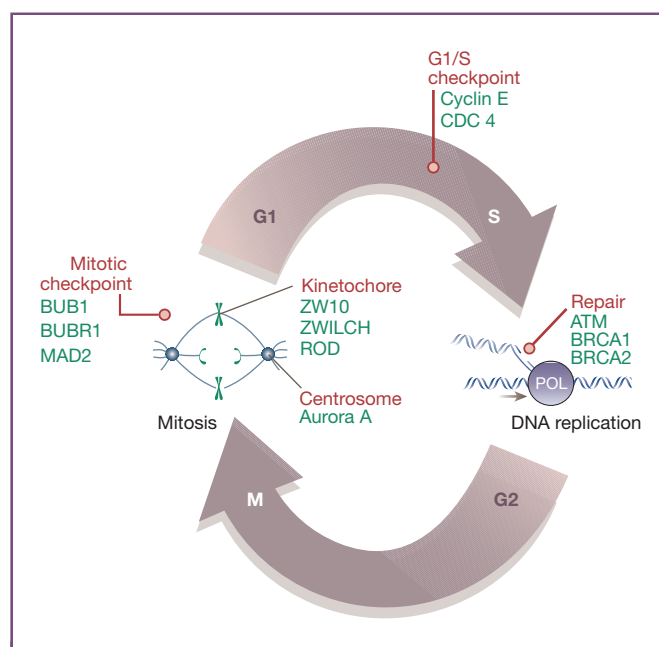


Figure 2 Multiple roads to aneuploidy. The schematic illustrates a simplified cell cycle, highlighting processes that have been implicated in the advent of aneuploidy. Several pathways within the cell cycle (indicated in red) can be disrupted. Genes (indicated in green) associated with these processes and structures have been found to be mutated or functionally altered in aneuploid cancers.

processes have been reviewed elsewhere^{27–29}, but we shall focus on those that meet one or both of two specific criteria. A candidate can be called a 'CIN gene' only if: (1) there is unambiguous evidence for an alteration in an encoded protein's expression or function in cancers compared to normal cells; and (2) that reconstitution of this alteration confers CIN on a diploid cell in tissue culture³⁰.

The first example of CIN being caused by mutations in specific genes came from studies of the mitotic-spindle-checkpoint genes *hBUB1* and *hBUBR1* (ref. 25). These genes are mutated infrequently in colorectal cancers and in other tumour types, and mutant alleles of *hBUB1* and reduction of murine BUBR1 expression confer CIN on dividing diploid cells^{25,31,32}. Recently, hereditary mutations in *hBUBR1* have been found in individuals with mosaic variegated aneuploidy, a rare disease with increased cancer risk³³. Another spindle component gene, *hMAD2*, is transcriptionally silenced in breast and other cancers³⁴. Genetic inactivation of a single allele of *hMAD2* or overexpression (perhaps as a result of inactivation of the retinoblastoma (Rb) tumour suppressor pathway) confers CIN on diploid cells^{35,36}. Mutations in other kinetochore- and centromere-associated genes have also been identified, such as *hZW10*, *hZWILCH*, and *hROD*, but evidence for their role in CIN so far stems purely from their homology to CIN genes in *Drosophila* or yeast³⁷.

Numerous other proteins function upstream of the kinetochore but are nevertheless integral to chromosome segregation, and may therefore play a role in CIN. In addition to kinetochore components and spindle-checkpoint proteins, the centrosome has been implicated in chromosomal instability. So far, no mutations in any genes encoding centrosome components have been identified in human cancer. Amplification and overexpression of Aurora-A kinase (STK15/BTAK) has been observed in tumours in association with centrosome amplification and aneuploidy³⁸.

Even further upstream, amplification/overexpression of *cyclin E* and mutation of *hCDC4* (both previously implicated in G1–S phase transitions) have been identified in aneuploid cancers and have also been associated with CIN (refs 39, 40). Intriguingly, cells lacking functional *hCDC4* demonstrate the same inability to arrest in mitosis in response to microtubule-disrupting agents as cells with mutations in kinetochore genes (D. A. Dezentje and S. Kern, personal communication). Therefore, it seems that even mutations and alterations of genes that might be expected to act at points temporally outside the metaphase–anaphase transition can lead to failure of the spindle checkpoint in CIN cancers²⁶.

In addition, hereditary mutations in recombination and DNA-repair genes (*ATM*, *BRCA1* and *BRCA2*) that have been shown to cause tumour predisposition presumably act by initiating CIN. Mouse knockout experiments have provided functional evidence to support the idea that these genes can contribute to aneuploidy^{41,42}. However, whether mutational inactivation of these genes results in increased rates of chromosomal gains/losses has not been elucidated.

It is perhaps not surprising that several routes leading to chromosomal instability, rather than a single mechanism, are now being uncovered in cancer cells. In yeast, mutations in over 100 genes can lead to CIN (refs 43, 44). These genes normally express proteins that have roles in disparate aspects of cellular life and would not otherwise be linked. Ultimately, our understanding of CIN will require the same comprehensive genetic and cell biological analysis of human genes and proteins that has already been accomplished in unicellular organisms. In the past few years we have witnessed an explosion in this sort of phenotypic dissection, and are hopeful that soon we will develop the level of understanding necessary to account for most genetic causes of aneuploidy in all cancers.

Implications for cancer prognosis and therapy

The scientific pursuit of aneuploidy and chromosomal instability is a fascinating biological question, but more significantly, it has the potential for tremendous clinical impact. First, aneuploid and/or chromosomally unstable cancers are likely to have a poorer prognosis

than diploid cancers — the degree of aneuploidy correlating with the severity of the disease^{45,46}. The reasons for the poorer prognosis is unclear, and somewhat surprising given that mismatch repair deficient cancers would otherwise be expected to have greater genetic fitness than chromosomally unstable cancers⁴. Another important point about prognosis is that ploidy status is generally not used to guide treatment regimes for patients with solid tumours. This is true despite the fact that ploidy seems more predictive of clinical outcome than other commonly used tools⁴⁷.

Chromosomal instability may also contribute to a cancer's ability to acquire chemoresistance. Although mechanisms of resistance to common chemotherapeutic agents are largely unknown (see review in this issue by Lowe *et al.*, page 307), a few select cases indicate that chromosomal instability may have an important role. For instance, resistance to Gleevec (a tyrosine kinase inhibitor, also known as STI-571), which is used in the treatment of chronic myelogenous leukaemia (CML) and a few other tumour types, has been shown to arise in CML patients whose cancers have undergone blast crisis⁴⁸ (see also introduction in this issue by Sawyers, page 294). Resistance to Gleevec arises from point mutations within the BCR–ABL fusion protein that is the hallmark of CML, but also from amplifications and chromosome duplications in chromosomally unstable CML cells. A second example can be found in the amplification of the thymidylate synthase gene in aneuploid cancers resistant to 5-fluorouracil⁴⁹. Both results are intriguing because they show that a change in the micro-environment can result in an amazingly rapid replacement of all tumour cells in the population with the progeny of a variant cell that was innately resistant to the drug. This provides a dramatic illustration of the clonal evolution of tumours and a cogent lesson about how powerful CIN can be in this process. CIN probably contributes to tumour evolution in similar ways in the absence of drugs, facilitating the generation of occasional cells that have the capacity to grow more efficiently in hostile environments set up by natural host defence mechanisms.

Both the prognostic relevance of chromosomal instability and its apparent capacity to allow evasion of chemotherapeutic intervention beg for the development of rationalized therapy targeting CIN. There are at least two ways such intervention could work. One method would involve targeting chromosomal instability directly. Recent elegant experiments provide proof of principle for such a strategy with implications for inhibiting proliferation of tumour cells: reducing the concentrations of the checkpoint proteins BUBR1 or MAD2 or inhibiting BUBR1 kinase activity can provoke apoptotic cell death in aneuploid human cancer cells. Thus, suppression of mitotic-checkpoint signalling seems invariably lethal as the consequence of massive chromosome loss⁵⁰. Both the BUB and MAD gene families were originally identified in yeast because mutations in these genes conferred resistance to mitotic arrest in response to particular pharmacological agents⁴³. It is possible that anti-CIN drugs can be found by doing the exact reverse. Starting with genetic causes of chromosomal instability, high-throughput drug screens for small molecule inhibitors of chromosomal instability that act in a genotype-specific manner could be found.

Another method to target CIN would be to find agents that inhibit pathways necessary to maintain chromosomal instability. In this way, anti-CIN compounds could serve as adjunct chemotherapy, preventing cancers from acquiring the mutations that allow them to develop resistance to other cytotoxic agents.

Taking these mechanistic and clinical insights together, we anticipate that the key challenges in the next ten years will be to elucidate the complicated mechanisms underlying aneuploidy and genomic instability, and to creatively apply this knowledge to cancer therapy. We are optimistic that the role of different cellular processes in aneuploidy will lead to concerted efforts to develop compounds that will target this feature common to most, if not all, cancers. The time has come to start applying our understanding of these processes in a clinically meaningful way. If we answer this call, aneuploidy will

become known not only as one of the oldest recognized properties of cancer, but as a property integral to the development of new therapies for this disease. □

doi:10.1038/nature03099

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
2. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
3. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nature Rev. Cancer* **1**, 157–162 (2001).
4. Nowak, M. A. *et al.* The role of chromosomal instability in tumor initiation. *Proc. Natl Acad. Sci. USA* **99**, 16226–16231 (2002).
5. Hanseemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Virchows Arch. Pathol. Anat.* **119**, 299–326 (1890).
6. Boveri, T. *Zur Frage der Entstehung maligner Tumoren* (Gustav Fischer Verlag, Jena, 1914).
7. Duesberg, P. *et al.* How aneuploidy may cause cancer and genetic instability. *Anticancer Res.* **19**, 4887–4906 (1999).
8. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
9. Hanseemann, D. Ueber pathologische Mitosen. *Arch. Pathol. Anat. Phys. Klin. Med.* **119**, 299–326 (1891).
10. Saunders, W. S. *et al.* Chromosomal instability and cytoskeletal defects in oral cancer cells. *Proc. Natl Acad. Sci. USA* **97**, 303–308 (2000).
11. Gisselsson, D. *et al.* Centrosomal abnormalities, multipolar mitoses, and chromosomal instability in head and neck tumours with dysfunctional telomeres. *Br. J. Cancer* **87**, 202–207 (2002).
12. Dutrillaux, B., Gerbault-Seureau, M., Remvikos, Y., Zafrani, B. & Prieur, M. Breast cancer genetic evolution: I. Data from cytogenetics and DNA content. *Breast Cancer Res. Treat.* **19**, 245–255 (1991).
13. Reid, B. J., Haggitt, R. C., Rubin, C. E. & Rabinovitch, P. S. Barrett's esophagus: correlation between flow cytometry and histology in detection of patients at risk for adenocarcinoma. *Gastroenterology* **93**, 1–11 (1987).
14. Levine, D. S. *et al.* Distribution of aneuploid cell populations in ulcerative colitis with dysplasia or cancer. *Gastroenterology* **101**, 1198–1210 (1991).
15. der-Sarkissian, H., Bacchetti, S., Cazes, L. & Londono-Vallejo, J. A. The shortest telomeres drive karyotype evolution in transformed cells. *Oncogene* **23**, 1221–1228 (2004).
16. Meraldi, P., Honda, R. & Nigg, E. A. Aurora-A overexpression reveals tetraploidization as a major route to centrosome amplification in p53^{-/-} cells. *EMBO J.* **21**, 483–492 (2002).
17. Lingle, W. L. *et al.* Centrosome amplification drives chromosomal instability in breast tumor development. *Proc. Natl Acad. Sci. USA* **99**, 1978–1983 (2002).
18. Pihan, G. A. *et al.* Centrosome defects and genetic instability in malignant tumors. *Cancer Res.* **58**, 3974–3985 (1998).
19. Lynch, H. T. & de la Chapelle, A. Hereditary colorectal cancer. *N. Engl. J. Med.* **348**, 919–932 (2003).
20. Marra, G. & Boland, C. R. Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. *J. Natl Cancer Inst.* **87**, 1114–1125 (1995).
21. Peltomäki, P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Hum. Mol. Genet.* **10**, 735–740 (2001).
22. Yamamoto, H., Imai, K. & Perucho, M. Gastrointestinal cancer of the microsatellite mutator phenotype pathway. *J. Gastroenterol.* **37**, 153–163 (2002).
23. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623–627 (1997).
24. Heim, S. & Mitelman, F. in *Cancer Cytogenetics* (eds Heim, S. & Mitelman, F.) 369–388 (Wiley-Liss, New York, 1995).
25. Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
26. Tighe, A., Johnson, V. L., Albertella, M. & Taylor, S. S. Aneuploid colon cancer cells have a robust spindle checkpoint. *EMBO Rep.* **2**, 609–614 (2001).
27. Hoyt, M. A. A new view of the spindle checkpoint. *J. Cell Biol.* **154**, 909–912 (2001).
28. Fukagawa, T. Assembly of kinetochores in vertebrate cells. *Exp. Cell Res.* **296**, 21–27 (2004).
29. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
30. Rajagopalan, H., Nowak, M. A., Vogelstein, B. & Lengauer, C. The significance of unstable chromosomes in colorectal cancer. *Nature Rev. Cancer* **3**, 695–701 (2003).
31. Bardelli, A. *et al.* Carcinogen-specific induction of genetic instability. *Proc. Natl Acad. Sci. USA* **98**, 5770–5775 (2001).
32. Dai, W. *et al.* Slippage of mitotic arrest and enhanced tumor development in mice with BubR1 haploinsufficiency. *Cancer Res.* **64**, 440–445 (2004).
33. Hanks, S. *et al.* Constitutional aneuploidy and cancer predisposition caused by biallelic *BUB1B* mutations. *Nature Genet.* advance online publication 10 October 2004 (doi:10.1038/ng1449).
34. Li, Y., Gorbea, C., Mahaffey, D., Rechsteiner, M. & Benezra, R. MAD2 associates with the cyclosome/anaphase-promoting complex and inhibits its activity. *Proc. Natl Acad. Sci. USA* **94**, 12431–12436 (1997).
35. Michel, L. S. *et al.* MAD2 haplo-insufficiency causes premature anaphase and chromosome instability in mammalian cells. *Nature* **409**, 355–359 (2001).
36. Hernando, E. *et al.* Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. *Nature* **430**, 797–802 (2004).
37. Wang, Z. *et al.* Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer Res.* **64**, 2998–3001 (2004).
38. Zhou, H. *et al.* Tumour amplified kinase STK15/BTK induces centrosome amplification, aneuploidy and transformation. *Nature Genet.* **20**, 189–193 (1998).
39. Strohmaier, H. *et al.* Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line. *Nature* **413**, 316–322 (2001).
40. Rajagopalan, H. *et al.* Inactivation of hCdc4 can cause chromosomal instability. *Nature* **428**, 77–81 (2004).
41. Weaver, Z. *et al.* Mammary tumors in mice conditionally mutant for Brca1 exhibit gross genomic instability and centrosome amplification yet display a recurring distribution of genomic imbalances that is similar to human breast cancer. *Oncogene* **21**, 5097–5107 (2002).
42. Tutt, A. *et al.* Absence of Brca2 causes genome instability by chromosome breakage and loss associated with centrosome amplification. *Curr. Biol.* **9**, 1107–1110 (1999).
43. Spencer, F., Gerring, S. L., Connelly, C. & Hieter, P. Mitotic chromosome transmission fidelity mutants in *Saccharomyces cerevisiae*. *Genetics* **124**, 237–249 (1990).
44. Hieter, P., Mann, C., Snyder, M. & Davis, R. W. Mitotic stability of yeast chromosomes: a colony color assay that measures nondisjunction and chromosome loss. *Cell* **40**, 381–392 (1985).
45. Zhou, W. *et al.* Counting alleles to predict recurrence of early-stage colorectal cancers. *Lancet* **359**, 219–225 (2002).
46. Watanabe, T. *et al.* Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N. Engl. J. Med.* **344**, 1196–1206 (2001).
47. Risques, R. A. *et al.* Genetic pathways and genome-wide determinants of clinical outcome in colorectal cancer. *Cancer Res.* **63**, 7206–7214 (2003).
48. Sawyers, C. L. Research on resistance to cancer drug Gleevec. *Science* **294**, 1834 (2001).
49. Wang, T. L. *et al.* Digital karyotyping identifies thymidylate synthase amplification as a mechanism of resistance to 5-fluorouracil in metastatic colorectal cancer patients. *Proc. Natl Acad. Sci. USA* **101**, 3089–3094 (2004).
50. Kops, G. J., Foltz, D. R. & Cleveland, D. W. Lethality to human cancer cells through massive chromosome loss by inhibition of the mitotic checkpoint. *Proc. Natl Acad. Sci. USA* **101**, 8699–8704 (2004).

Acknowledgments We thank the members of our laboratory for critical reading of this manuscript. We also thank D. Gisselsson and M. Höglund for sharing some of their unpublished thoughts on mitotic disturbances, and M. R. Speicher for providing the photograph for Fig. 1. We especially thank S. Kern and D. A. Dezentje for permission to report unpublished data.

Competing interests statement The authors declare that they have no competing financial interests.

Endurance running and the evolution of *Homo*

Dennis M. Bramble¹ & Daniel E. Lieberman²

¹Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA

²Peabody Museum, Harvard University, Cambridge, Massachusetts 02138, USA

Striding bipedalism is a key derived behaviour of hominids that possibly originated soon after the divergence of the chimpanzee and human lineages. Although bipedal gaits include walking and running, running is generally considered to have played no major role in human evolution because humans, like apes, are poor sprinters compared to most quadrupeds. Here we assess how well humans perform at sustained long-distance running, and review the physiological and anatomical bases of endurance running capabilities in humans and other mammals. Judged by several criteria, humans perform remarkably well at endurance running, thanks to a diverse array of features, many of which leave traces in the skeleton. The fossil evidence of these features suggests that endurance running is a derived capability of the genus *Homo*, originating about 2 million years ago, and may have been instrumental in the evolution of the human body form.

Most research on the evolution of human locomotion has focused on walking. There are a few indications that the earliest-known hominids were bipeds^{1,2}, and there is abundant fossil evidence that australopithecines habitually walked by at least 4.4 million years (Myr) ago^{3,4}. Many researchers interpret the evolution of an essentially modern human-like body shape, first apparent in early *Homo erectus*, as evidence for improved walking performance in more open habitats that came at the expense of retained adaptations in the australopithecine postcranium for arboreal locomotion (for example, refs 5–8). Although the biomechanics of running, the other human gait, is well studied, only a few researchers (see refs 9, 10 for example) have considered whether running was a mode of locomotion that influenced human evolution. This lack of attention is largely because humans are mediocre runners in several respects. Even elite human sprinters are comparatively slow, capable of sustaining maximum speeds of only 10.2 m s⁻¹ for less than 15 s. In contrast, mammalian cursorial specialists such as horses, greyhounds and pronghorn antelopes can maintain maximum galloping speeds of 15–20 m s⁻¹ for several minutes¹¹. Moreover, running is more costly for humans than for most mammals, demanding roughly twice as much metabolic energy per distance travelled than is typical for a mammal of equal body mass¹². Finally, human runners are less manoeuvrable and lack many structural modifications characteristic of most quadrupedal cursors such as elongate digitigrade feet and short proximal limb segments.

However, although humans are comparatively poor sprinters, they also engage in a different type of running, endurance running (ER), defined as running many kilometres over extended time periods using aerobic metabolism. Although not extensively studied in non-humans, ER is unique to humans among primates, and uncommon among quadrupedal mammals other than social carnivores (such as dogs and hyenas) and migratory ungulates (such as wildebeest and horses)^{13,14}. Here, we review the evidence for and impact of ER in human evolution. We begin with a discussion of the mechanical differences between walking and running, and how well humans perform at ER compared to other mammals. We then review what is known about the key structural specializations thought to underlie human ER capabilities, the extent to which they may be features that evolved originally for bipedal walking, and the evidence for their appearance in the hominid fossil record. We conclude by outlining some hypotheses for why ER capabilities initially arose in the genus *Homo*, and the significance of this behaviour for human evolution.

How well do humans run long distances?

In considering human running, it helps to start from the perspective of the basic biomechanical differences that distinguish running and walking gaits in all mammals, including human bipeds. These differences are well characterized. Walking uses an ‘inverted pendulum’ in which the centre of mass vaults over a relatively extended leg during the stance phase, efficiently exchanging potential and kinetic energy out-of-phase with every step (Fig. 1a, b). The metabolic cost of transport (COT) for human walking, like that of other mammals, is a ‘U’-shaped curve, in which optimal speed, approximately 1.3 m s⁻¹, is largely a function of leg length¹⁵. Most humans voluntarily switch to running at approximately 2.3–2.5 m s⁻¹, which corresponds closely to the intersection of the COT curves for walking and running in humans (Fig. 2b)^{16,17}. At these higher speeds running becomes less costly than walking by exploiting a mass-spring mechanism that exchanges kinetic and potential energy very differently (Fig. 1b). Collagen-rich tendons and ligaments in the leg store elastic strain energy during the initial, braking part of the support phase, and then release the energy through recoil during the subsequent propulsive phase^{18,19}. To use these springs effectively, the legs flex more in running than in walking: flexing and then extending at the knee and ankle during the support phase (Fig. 1a). Limb stiffness relative to body mass in running humans is similar to that of other mammalian cursors²⁰.

Although extensive data on endurance capabilities are not available for most quadrupedal mammals, several lines of evidence indicate that humans, using criteria such as speed and sustainable distance, are much better endurance runners than has generally been appreciated. Human ER speeds range from approximately 2.3 to as much as 6.5 m s⁻¹ in elite athletes. Average ER speeds for recreational joggers range between 3.2–4.2 m s⁻¹ (ref. 21). From an evolutionary perspective, it is important to note that human ER speeds are exceptional compared to non-human primates. Apes such as chimpanzees, and other primates, such as patas monkeys, can sprint rapidly, but they do so rarely and only for short distances^{22,23}. No primates other than humans are capable of ER.

Quadrupedal cursors easily sprint faster than humans over short distances, but sustainable ER speeds of humans are surprisingly comparable to specialized mammalian cursors such as dogs and horses in two respects. The first comparison to make is with trotting, because bipeds are incapable of galloping, but also because human bipedal running and quadrupedal trotting are biomechanically most comparable. Both gaits synchronize contralateral fore- and hindlimbs, effectively restricting each stride cycle to just two steps, and both are inherently ‘bouncy’ gaits with substantial vertical displacements of the centre of mass^{18,24}. When compared

to quadrupedal trotting, human ER speeds are relatively high when adjusted for body mass (Fig. 2a). The predicted preferred trotting speed for a human-sized (65 kg) quadruped is approximately 2.8 m s^{-1} , and the trot–gallop transition is 3.8 m s^{-1} (ref. 25). A more extreme comparison of performance that is not adjusted for body size is between humans and large mammals such as ponies and horses (Fig. 2a). Human ER speeds exceed the preferred trotting (3.1 m s^{-1}) and the trot–gallop transition (4.4 m s^{-1}) speeds of ponies (110–170 kg)²⁶, and even the preferred trotting speed predicted for a 500-kg quadruped²⁵.

Most cursorial quadrupeds such as zebra, antelopes, and African hunting dogs trot when running long distances¹⁴, but a few such as hyenas and wildebeest are known to run long distances using a low-speed gallop (typically a canter)¹³. When galloping, species with high sustainable speeds such as dogs or horses can usually outrun humans. The maximum sustainable (~ 10 – 15 min) galloping speed predicted for a 65-kg quadruped is 7.7 m s^{-1} , and elite racing horses can gallop 10 km at 8.9 m s^{-1} (refs 25, 27). However, human ER speeds are quite comparable to the preferred galloping speeds that cursors use over longer distances and times. Minetti²⁷ has shown that sustainable galloping speeds in horses decline considerably for runs longer than 10–15 min, accounting for the average daytime speed of 5.8 m s^{-1} at which long-distance postal horses were consistently run for millennia. Wildebeests (~ 100 kg) prefer to canter at 5.1 m s^{-1} (ref. 13). Well-conditioned human runners exceed the predicted preferred galloping speed for a 65-kg quadruped²⁵ and can occasionally outrun horses over the extremely long distances that constrain these animals to optimal galloping speeds, typically a canter (Fig. 2a)^{9,10}.

Humans also perform well at ER by another criterion, sustainable distance. Approximately 10% of Americans habitually jog or run several kilometres a day (the percentage is higher if one includes treadmill exercise and related sports)²⁸. Fit human amateurs can regularly run 10 km, and longer distances such as marathons (42.2 km) are achieved by tens of thousands of people each year. Such distances are unknown if not impossible for any other primate, but are comparable to those observed in specialized mammalian cursors in open habitats. African hunting dogs travel an average of 10 km per day, and wolves and hyenas travel on average 14 and 19 km day⁻¹, respectively¹⁴. This is not to say that humans can

outdistance specialized quadrupeds. Some horse and dog breeds, for example, can be made to run more than 100 km day^{-1} while carrying or pulling a human. Such extreme and human-induced feats, however, should not detract from the fact that humans can and do run long distances well, despite a primate ancestry.

The one category in which humans perform poorly compared to many quadrupeds is the energetic cost of running. The mass-adjusted COT of human running is about 50% higher than a typical mammal, including other primates¹². Compared to the only value measured for a chimpanzee (a 17.5-kg juvenile), human running is 25% less costly in absolute terms, but about 10% more costly when adjusted for body mass²⁹. Interestingly, other endurance cursors such as wolves and African hunting dogs also have high mass-adjusted COT relative to the average mammal¹². One important characteristic of human ER may be its range of accessible economical speeds. Horses have U-shaped COT curves with narrow ranges of preferred speeds for trotting and galloping and gait transitions that minimize cost, thereby achieving an effectively flat COT curve that excludes many speeds within the aerobic range (Fig. 2b)²⁶. It is not known whether other quadrupedal cursors such as dogs have U-shaped COT curves, but human runners differ from horses in employing a single gait, with a flat COT curve at all but the fastest endurance speeds^{9,16}. Like another group of cursorial bipeds, kangaroos and wallabies, humans are thus able to adjust running speed continuously without change of gait or metabolic penalty over a wide range of speeds. Further research is necessary to determine whether other cursors are capable of such a broad range of economic speeds.

Structural bases and fossil evidence for endurance running

The human capacity for ER raises several questions. What features make ER possible? When do these features first appear in the fossil record? How might such features relate to adaptations for bipedal walking? Many of the anatomical and physiological features involved in running are well studied in mammals, including humans, but most have not been explicitly evaluated in the human fossil record. A useful approach is to consider separately the evidence for structural features relevant to four types of demands posed by ER: energetics, strength, stabilization and thermoregulation. The skeletal traces of these features, and the

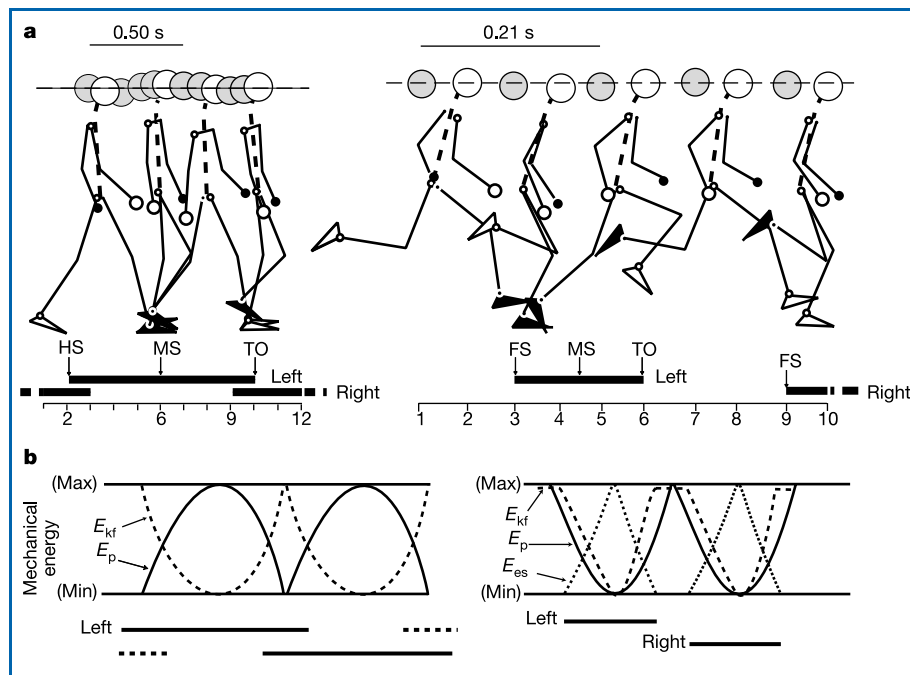


Figure 1 Comparisons of walking and running.

a, Kinematics of walking (left) and running (right), from plates 4 and 18 of ref. 64. During walking, the head and centre of gravity are lowest near toe-off (TO) and highest at mid-stance (MS) where the leg is relatively straight. During running, the head and centre of gravity are highest during the aerial phase and lowest at MS, when the hip, knee and ankle are flexed; the trunk is also more inclined and the elbow more flexed.

b, Biomechanical contrasts between human gaits. During walking, an inverted pendulum mechanism exchanges forward kinetic energy (E_k) for gravitational potential energy (E_p) between heelstrike (HS) and MS; the exchange is reversed between MS and TO. During running, a mass-spring mechanism causes E_p and E_k to be in phase, with both energies declining rapidly to minima between footstrike (FS) and MS. Leg tendons and ligaments partially convert decreases in E_p and E_k to elastic strain energy (E_{es}) during the first half of the stance, which is subsequently released through recoil between MS and TO.

evidence for their first presence in the fossil record, are summarized in Table 1 and illustrated in Fig. 3. Several issues need to be kept in mind when evaluating these features.

First, it is useful to distinguish between structures that benefit both walking and running from those that are specific to the unique biomechanics of running and are functionally unrelated to walking. Second, the limitations of the fossil record complicate our ability to test evolutionary hypotheses concerning many structural modifications that are derived in humans relative to chimpanzees. Some, such as Achilles tendon length, leave no clear skeletal evidence—rendering uncertain their first appearance. Others, particularly in the foot, are not yet adequately sampled in the fossil record to make it possible to identify their origins.

Energetics

Humans exhibit many musculoskeletal specializations for bipedalism. Given the fundamental biomechanical contrasts between walking and running, which features are specifically relevant to the energetic cost of running? As noted above, the mass-spring mechanics of running differ from the pendular mechanics of walking: running uses a compliant limb in which muscles and tendons in the legs sequentially store and then release strain energy during the stance phase of the stride cycle. In contrast to apes, human legs have many long spring-like tendons connected to short muscle fascicles that can generate force economically³⁰. These springs (see Fig. 3) can have comparatively little effect on energy savings during an inverted pendulum-like walk, particularly at heel strike when the limb is not compliant, but are estimated to save approximately 50% of the metabolic cost of running^{17,19}. The most important of these springs is the Achilles tendon, which connects the heel with the major plantar flexors of the foot; other elongated tendons that are derived features of the human leg include the iliotibial tract and m. (muscle) peroneus longus³¹. Unfortunately, there are no preserved early *Homo* calcanei, and leg tendon length probably cannot be estimated reliably from attachment sites.

However, the transverse groove into which the Achilles tendon inserts on the posterior surface of the calcaneus is chimpanzee-like in size in three early australopithecine Hadar specimens (AL 333-8, 333-37 and 333-55)^{32,33}, and contrasts with the substantially wider and taller attachment area characteristic of *H. sapiens*. We hypothesize that, as in modern apes, a developed Achilles tendon was absent in *Australopithecus* and originated at some point after 3 Myr ago, probably in the genus *Homo*.

Another well-developed set of springs important to human running is the longitudinal arch of the foot. During walking, the plantar arch helps to maintain mid-tarsal rigidity for powered plantar flexion during toe-off, and absorbs some impact force (but only after heel strike); during running, the elastic structures of the plantar arch function as a spring, returning approximately 17% of the energy generated during each stance phase¹⁹. Several features in australopithecine foot bones from Hadar and Sterkfontein (STW 573) suggest that some sort of plantar arch was present, including an elongated lateral cuneiform and insertions for the plantar ligaments^{4,34,35}. But analyses of the Hadar and Sterkfontein specimens suggest that they may have had a partial arch only, as indicated by the enlarged medial tuberosity of the navicular, which is also enlarged and weight-bearing in chimpanzees, but is diminutive and not weight-bearing in *Homo*³⁶. In addition, for the plantar arch to be an effective spring during running, the transverse tarsal joint must restrict rotation between the hind foot and the anterior tarsals, allowing passive stretching of the plantar ligaments during a mid-foot strike. In humans, this rotation is restricted primarily by a projecting medial flange on the proximal cuboid, which causes the calcaneocuboid joint to form a close-packed position following several degrees of rotation³⁷. There are no preserved early *H. erectus* feet, but this feature—together with a fully adducted big toe—is first apparent in the OH 8 foot^{4,36,37}, which is generally ascribed to *H. habilis*.

An additional energetic factor to consider is stride length. Unlike most quadrupeds²⁵, humans increase speed during ER mostly by

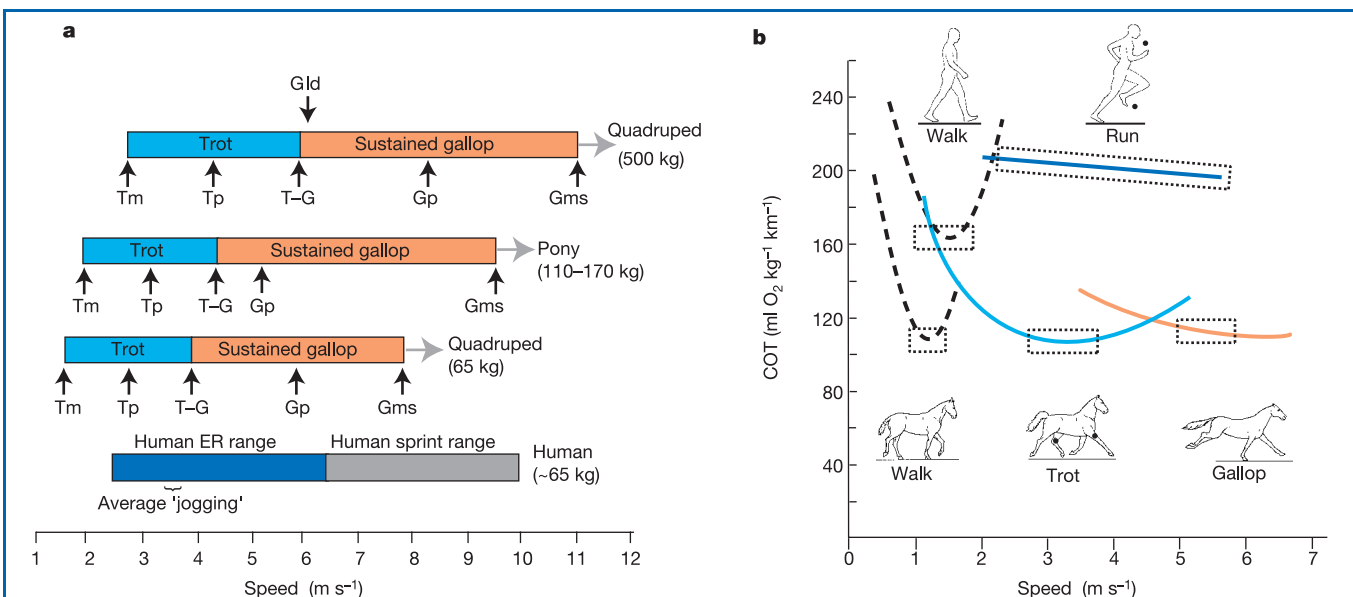


Figure 2 Comparative ER performance in humans and quadrupeds. **a**, Range of speeds for human ER and sprinting, and minimum trot (Tm), preferred trot (Tp), trot-gallop transition (T-G), preferred gallop (Gp), and maximum sustained gallop (Gms) for ponies (ref. 26), and predicted for quadrupeds of 65 and 500 kg (ref. 25). Also indicated is Gld, the optimal long distance (~20 km), daytime galloping speed for horses (ref. 27). Note that quadrupeds sprint at speeds above Gms. **b**, Comparison of the metabolic cost of transport (COT) in humans and ponies^{9,16,17}. Both species

have U-shaped COT curves for walking, and trotting has a similar-shaped curve in the horse, but the human COT is essentially flat at ER speeds. Preferred speeds (dotted rectangles) correspond to the most energy-efficient speeds in horses and walking humans, but speed selection is unrestricted in human ER. Note also that human running, like quadrupedal trotting, involves synchronized movements of diagonally opposite appendages (dots).

increasing stride length rather than rate (Fig. 4). Stride lengths in humans during ER are typically more than 2 m, and can exceed 3.5 m in elite runners²¹, approximately a metre longer than the strides predicted for a 65-kg quadruped²⁵ or measured in chimpanzees³⁸ at the same speeds, even when galloping (Fig. 4a). Long absolute (rather than relative) stride lengths in humans are made possible by a combination of effective leg springs (see above) and relatively long legs. Long legs benefit walking by increasing optimum walking speed, but they also increase ground contact time in both walking and running¹⁵. Relatively long contact times may be advantageous for ER because the inverse of contact time has been found to correlate across species with the energetic cost of running (running is priced by the step)³⁹. Long legs relative to body mass, typical of most specialized cursors, first appear unequivocally in hominids 1.8 Myr ago with *H. erectus*, whose relative leg length (assessed from the femur) is possibly up to 50% greater than in *Australopithecus afarensis*⁴. Leg length in *H. habilis* (estimated from the OH 62 skeleton) and other specimens as early as 2.5 Myr ago is currently the subject of debate⁴⁰.

Oscillating long legs, however, increases the energy cost of running in proportion to the limb's mass moment of inertia. Reductions in distal limb mass have little effect on the energetics of walking but produce substantial metabolic savings during ER, roughly proportional to the square of the distance of the mass from the hip. Redistributing 3.6 kg from the ankles to the hip, for example, decreases the metabolic cost of human running at slow speeds (2.6 m s^{-1}) by 15% (ref. 41). Although we do not know the relative mass of the distal limb in fossil hominids, humans differ from australopithecines^{4,32}, and resemble many specialized cursors in having more compact feet and relatively short toes; the human foot is only 9% of total leg mass, compared to 14% in chimpanzees⁴². Humans also have relatively low stride rates at ER speeds, even lower than are predicted for a 500-kg quadruped²⁵ (Fig. 4b). Low stride rates that increase little in the ER range reduce the force required to oscillate the heavy legs (30% of body mass in humans, compared to 18% in chimpanzees⁴²) and may favour greater

reliance on more slowly contracting, oxidative and fatigue-resistant muscle fibres, which are relatively more abundant in the legs of competitive distance runners than in sprinters⁴³. The high percentage of slow-twitch muscle fibres necessary for endurance running may have originated in humans from a novel null mutation of the *ACTN3* gene⁴⁴.

Skeletal strength

Another factor to consider when evaluating the evolution of ER in humans is skeletal strength. Running exposes the skeletal system to much higher stresses than walking, especially when the foot collides with the ground, producing a shock wave that passes up the body from the heel through the spine to the head. Peak vertical ground reaction forces (GRFs) at heel strike are approximately twice as high during running than during walking and may approach 3–4 times body weight at higher ER speeds⁴⁵. Human runners reduce these stresses to some extent through limb compliance and mid-foot striking (thereby also storing elastic strain energy in the leg and foot), but must otherwise dissipate impact forces within their bones and joints. One strategy to lower joint stress is to expand joint surfaces, spreading forces over larger areas. Many studies have found that compared to both *Pan* and *Australopithecus*, *Homo* has substantially larger articular surface areas relative to body mass in most joints of the lower body, including the femoral head and knee^{6,7}, the sacroiliac joint^{46,47}, and the lumbar centra⁴⁷. Enlargement of these joints, which is not matched in the upper limb of *Homo*⁶, lowers the stresses that impact forces generate at heel strike during walking, but would contribute more critically to dissipate the much higher impact loads generated in running. Another possible modification of the pelvis for resisting the stresses associated with running is enlargement of the iliac pillar in early *H. erectus*^{4,46}. Humans may also have a larger cross-sectional area of the calcaneal tuber relative to body mass than australopithecines³³.

Both walking and running also cause diaphyseal loading, which is higher in running and increases relative to body mass as a function

Table 1 Derived features of the human skeleton with cursorial functions

Feature	Functional role	W/R*	Earliest evidence
Enlarged posterior and anterior semicircular canals	Head/body stabilization	R	<i>H. erectus</i>
Expanded venous circulation of neurocranium	Thermoregulation	R > W	<i>H. erectus</i>
More balanced head	Head stabilization	R	<i>H. habilis</i>
Nuchal ligament (1)	Head stabilization	R	<i>H. habilis</i>
Short snout (2)	Head stabilization	R > W	<i>H. habilis</i>
Tall, narrow body form	Thermoregulation	R > W	<i>H. erectus</i>
Decoupled head and pectoral girdle (3)	Counter-rotation of trunk versus head	R	<i>H. erectus</i> ?
Low, wide shoulders (4)	Counter-rotation of trunk versus hips	R	<i>H. erectus</i> ?
Forearm shortening (5)	Counter-rotation of trunk	R	<i>H. erectus</i>
Narrow thorax (6)	Counter-rotation of trunk versus hips	R	<i>H. erectus</i> ?
Narrow and tall waist between iliac crest and ribcage (7)	Counter-rotation of trunk versus hips	R	<i>H. erectus</i> ?
Narrow pelvis (8)	Counter-rotation of trunk versus hips	R	<i>Homo</i> ?
Expanded lumbar centra surface area (9)	Stress reduction	R > W	
Enlarged iliac pillar (10)	Stress reduction	R > W	<i>H. erectus</i>
Stabilized sacroiliac joint	Stress reduction	R > W	<i>H. erectus</i>
Expanded surface area for mm. erector spinae origin (11)	Trunk stabilization	R	<i>H. erectus</i>
Expanded surface area for m. gluteus maximus origin (12)	Trunk stabilization	R	<i>H. erectus</i>
Long legs (13)	Stride length	R, W	<i>H. erectus</i>
Expanded hindlimb joint surface area (14)	Stress reduction	R > W	<i>H. erectus</i>
Shorter femoral neck (15)	Stress reduction	R > W	<i>H. sapiens</i>
Long Achilles tendon (16)	Energy storage	R	<i>Homo</i> ?
Plantar arch (passively stabilized) (17)	Shock absorption	R	<i>Homo</i> ?
	Energy storage	R	
	Shock absorption	R > W	
	Powered plantarflexion	R > W	
Enlarged tuber calcaneus (18)	Stress reduction	R > W	<i>Homo</i> ?
Close-packed calcaneocuboid joint	Energy storage	R	<i>H. habilis</i> (OH 8)
	Stability during plantarflexion	R > W	
Permanently adducted hallux (19)	Stability during plantarflexion	R > W	<i>H. habilis</i> (OH 8)
Short toes (20)	Stability during plantarflexion	R > W	<i>H. habilis</i> (OH 8)
	Distal mass reduction	R > W	

*W, R indicate traits that enhance performance in endurance walking and endurance running, respectively; R > W indicates traits that benefit both walking and ER, but which have a greater effect on ER. Numbers in parentheses correspond to those in Fig. 3a and c.

of speed⁴⁸. Like *Pan* and early *Homo*, australopithecines have robust femoral shafts relative to body mass, but they are less wide transversely than in early *Homo*⁷. Although the distinctly shorter femoral neck of humans compared to *Pan* or *Australopithecus* decreases the mechanical advantage of the hip abductors, it might also facilitate running by reducing bending moments in the femoral neck. The reduction in interacetabular hip breadth in *Homo* also reduces lateral bending moments on the pelvis and lower back generated at footstrike, and likewise helps minimize the angular momentum in the trunk caused by rapid oscillation of long, heavy legs⁴⁹.

Stabilization

Bipedal gaits are inherently unsteady, but several differences between running and walking call for special mechanisms during

running to help ensure stabilization and balance. Most obviously, the trunk and neck of human runners are more forwardly inclined during running than walking (Fig. 1a), resulting in a greater tendency to pitch forward, especially at heel strike. *Homo* has a number of derived features that enhance trunk stabilization, including expanded areas on the sacrum and the posterior iliac spine for the attachment of the large erector spinae muscles, and a greatly enlarged m. gluteus maximus^{4,46}. The latter muscle, whose increased size is among the most distinctive of all human features, is strongly recruited in running at all speeds but not in walking on level surfaces⁵⁰. In addition, the transverse processes of the sacrum are also relatively larger in *Homo* than *Australopithecus*, suggesting a more mechanically stable sacroiliac joint³⁴.

Independent rotations within the trunk play a crucial role in dynamic stabilization during human running and may help to

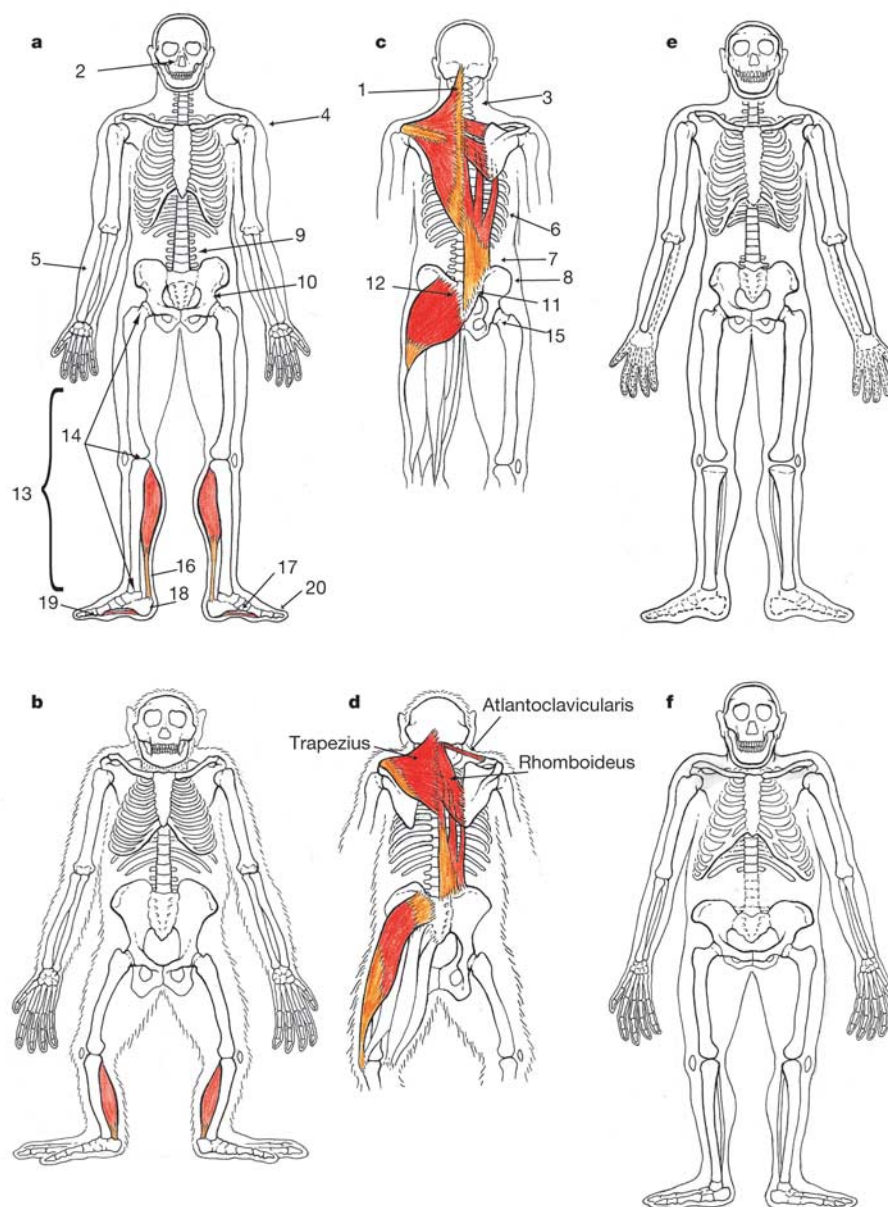


Figure 3 Anatomical comparisons of human, chimpanzee, *H. erectus* and *A. afarensis*. **a, c**, Anterior and posterior views of human, enumerating features related to endurance running listed in Table 1. **b, d**, Anterior and posterior views of chimpanzee. Labelled muscles connect the head and neck to the pectoral girdle and

are reduced or absent in humans. **e**, Reconstruction of *H. erectus* based primarily on KNM-WT 15000 (from refs 4, 65); **f**, reconstruction of *A. afarensis* based primarily on AL-288 (from refs 4, 66).

explain several derived features of *Homo*. In a walk, one leg is always on the ground, enabling the abductors and medial rotators of the stance hip to counteract the inertially induced rotation of the trunk (about its vertical axis) generated by the forward acceleration of the swing leg. However, during the aerial phase of running, leg acceleration generates even larger torques that cannot be counteracted by ground forces. These potentially destabilizing forces are offset by the opposing torques produced by counter-rotation of thorax and arms (but not the head)⁴⁹. At least three derived structural modifications in the hips and shoulder permit humans to generate these counter-balancing torques. First, humans are capable of a substantially greater degree of isolated rotation of the trunk relative to the hips compared to apes⁴, thanks to an elongate, narrow waist that vertically separates the lower margin of the thorax from the pelvis. This configuration is fully developed in *H. erectus*⁵¹. *Australopithecus* may have had a tall waist, but its broad, chimpanzee-shaped thorax and broad pelvis (possibly related to gut size⁵²) suggest a relatively wider waist than in *Homo*. Second, *Homo* differs from *Pan* and possibly from *Australopithecus* in having greater structural independence of the pectoral girdle and head. Chimpanzees have an inverted funnel-shaped upper thorax, with narrow and habitually elevated ('shrugged') shoulders, and extensive muscular connections (mm. (muscles) rhomboideus, atlanto-clavicularis, trapezius superior) between the shoulder and the head-neck complex that are either absent or much reduced in *Homo*^{4,31}. The cleidocranial portion of the m. trapezius is the sole muscular connection in humans between the pectoral girdle and head (Fig. 3c, d). Cranially oriented glenoid cavities (present in *Australopithecus*),

elevated shoulders and strong muscular connections to the head and neck are functionally advantageous for climbing³⁴, pose no obvious hindrance to bipedal walking, but would tend to impede the independent counter-rotations of the pectoral girdle and arms necessary to counter-balance the legs in running, and to minimize axial rotation of the head. (Decoupling of the head and pectoral girdle may also be advantageous for throwing.) Finally, the wide shoulders characteristic of *Homo* act to increase the counter-balancing moments generated by arm-swinging, while also permitting energy-saving reductions in forearm mass. Reductions in the forearm of *Homo* (50% less massive relative to total body mass in humans than chimpanzees^{4,42}), substantially lower the muscular effort required to maintain the stereotypically flexed elbow during ER.

Running also poses problems for head stabilization. Unlike quadrupeds, humans have vertically oriented necks that are less able to counteract the greater tendency of the head to pitch forward at foot strike during running than walking. Such inertial accelerations would be reduced in *Homo* relative to *Australopithecus* and *Pan* by a combination of decreased facial length and occipital projection behind the foramen magnum⁴. In addition, the radius of the posterior semicircular canal is significantly larger in *Homo* than in *Pan* or *Australopithecus*⁵³, presumably increasing the sensitivity of sensory perception to head pitching in the sagittal plane, which is potentially much greater during running than walking. Another possible structural modification relevant to running is the nuchal ligament, a convergent feature in *Homo* (first evident in KNM-ER 1813) and other mammals that are either cursorial (for example, dogs, horses, hares) or have massive heads (elephants)⁵⁴. Interestingly, a nuchal ligament is absent in chimpanzees^{4,31} and apparently in australopithecines (as evinced by the absence of a median nuchal line).

Thermoregulation and respiration

A final physiological challenge to consider is heat. Adaptations to maintain stable body temperature have long been considered important for long-distance walking in open, hot environments. However, running generates so much endogenous heat that sustained running is considerably more limited by thermoregulatory capabilities than is walking. As noted by refs 9 and 55, humans possess many derived features related to heat dissipation, including elaboration and multiplication of eccrine sweat glands for evapo-transpiration, and reduced body hair (which increases convection rates). We do not know when these non-musculoskeletal traits evolved, but several other derived features of *Homo* are possible mechanisms for dissipating metabolic heat, and could have been especially important for ER in hot environments. These include a narrow, elongated body form⁵⁶, and possibly an elaborated cranial venous circulation (for example, more accessory foramina in the cranial vault, and diploic expansion⁵⁷). The latter may use venous blood that has been cooled by sweating in the face and scalp to cool, via countercurrent heat exchange in the cavernous sinus, hot arterial blood in the internal carotid artery before it reaches the brain⁵⁸. Another derived feature of humans is the tendency for mouth breathing (but not panting) during strenuous activity. Nasal breathing, typical of apes, offers too much resistance within the relatively small human nasopharynx to support the high ventilatory demands of strenuous activities such as ER⁵⁹. Human distance runners are thus obligate mouth breathers, permitting higher airflow rates with less resistance and muscular effort; mouth breathing is also a more effective means of unloading excess heat during expiration.

Evolutionary hypotheses

Many hypotheses have been proposed for the role of walking (particularly long-distance trekking) in human evolution. Given human ER performance capabilities, as well as the many derived features that appear to make them possible, it is also necessary to ask

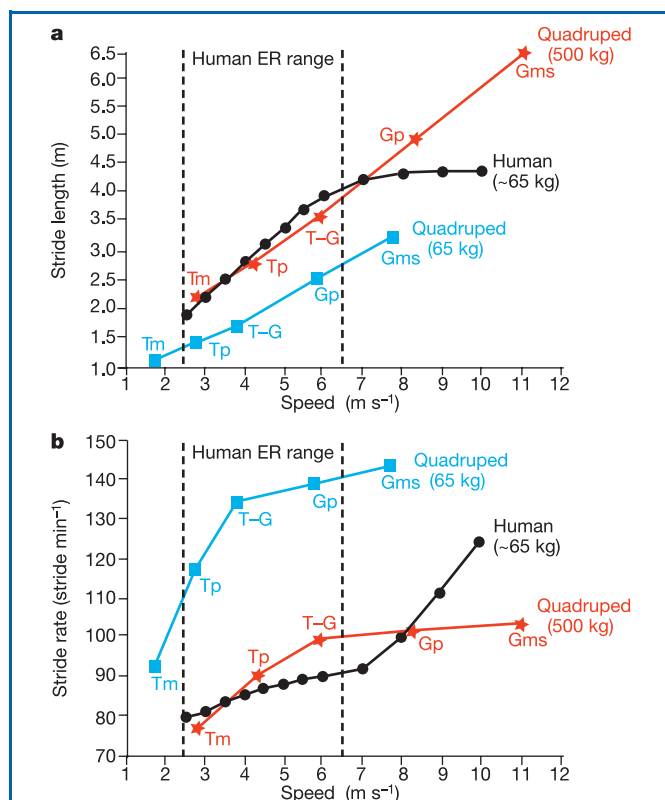


Figure 4 Comparison of stride length (a) and stride rate (b) contributions to running speed in humans^{21,64}, and in quadrupedal mammals (calculated from ref. 25) for various gaits (as in Fig. 2a). A stride is a complete locomotor cycle (two steps for a human). Compared to similar-sized quadrupeds, humans have relatively long stride lengths and relatively low stride rates in the ER range. Humans increase speed within the ER range primarily by increasing stride length not rate.

whether, when and why long-distance running may have played a role in human evolution. Although the fossil record is inadequate to pinpoint the origin of all the morphological features that contribute to human ER performance capabilities, most of the major structural bases of ER that can be observed in the skeleton are present in early *H. erectus* (Table 1). Despite disagreement over the hypodigm and systematic position of *H. habilis*⁸, several specimens that are generally attributed to this species (for example, the OH 8 foot and the KMN-ER 1813 cranium), also have a few derived features consistent with cursorial function (Table 1). It is thus reasonable to conclude that ER capabilities in human evolution originated in the genus *Homo*. Further data, however, are needed to test this hypothesis more fully. We currently lack any *H. erectus* feet, few postcranial remains are attributed to *H. habilis* or *H. rudolfensis*, and some key adaptations such as the length of the Achilles tendon are difficult and perhaps impossible to assess from fossils. Although the postcranial remains of australopithecines indicate that they walked habitually, their lack of any features associated with ER suggests that, like chimps³⁰, they probably did not run long distances well or frequently in the less-open habitats in which they lived.

The ER capabilities of *Homo* raise several additional questions, the first being whether long-distance running was an important behaviour in human evolution or merely the by-product of enhanced walking capabilities. Traditional arguments have favoured the latter hypothesis; several of the derived features of *Homo* in Table 1 are proposed as adaptations to improve long-distance walking performance in more arid, open habitats (for example, refs 5–8). These features include relatively longer legs, larger hindlimb and vertebral joint surfaces, narrower waists and shorter toes. Yet walking alone cannot account for many of the other derived features in Table 1 because the mass-spring mechanics of running, which differ fundamentally from the pendular mechanics of walking, require structural specializations for energy storage and stabilization that have little role in walking. Such specialized structures include: an extensive system of springs in the leg and foot that effectively store and release significant elastic energy during running; hypertrophied gluteus maximus and spinal extensor muscles that contract strongly to stabilize the trunk in running but not walking; and an elongate, narrow waist in combination with a low, wide, decoupled shoulder girdle that have an essential stabilizing function only in running.

Two additional lines of evidence suggest that ER capabilities in *Homo* are not solely by-products of selection for long-distance walking. First, sustained running poses extreme mechanical and thermoregulatory challenges beyond those encountered in distance walking. Expanded joint surfaces in the spine, hip, and legs, along with multiple specializations for shedding excess body heat (for example, sweating, hairlessness, cranial cooling systems), would be useful for prolonged walking in hot environments, but they would have been essential to tolerate the considerably higher impulsive loads and endogenous heat produced by distance running. Second, a few derived features of *Homo* that improve ER capabilities (notably forearm shortening and decoupling of the head and pectoral girdle) are unrelated to walking, but would have hindered arboreal locomotor capabilities. Thus some of the differences between *Homo* and *Australopithecus* that have been attributed to selection for more efficient long-distance walking may instead have evolved for ER, thereby helping to make *Homo* the first fully terrestrial hominoid.

Considering all the evidence together, it is reasonable to hypothesize that *Homo* evolved to travel long distances by both walking and running. New fossils and more detailed analyses of the existing fossil record are needed to test whether these two locomotor capabilities emerged concurrently or whether ER evolved after selection for long-distance walking. An even more difficult task is to determine what behaviours selected for ER in the first place. Why would early

Homo run long distances when walking is easier, safer and less costly? One possibility is that ER played a role in helping hominids exploit protein-rich resources such as meat, marrow and brain first evident in the archaeological record at approximately 2.6 Myr ago⁶⁰, coincident with the first appearance of *Homo*. Testing whether ER was employed in hunting or scavenging will be challenging given the limitations of the archaeological and ethnographic records. ER is not common among modern hunter-gatherers, who employ many technologies to hunt (for example, bows and arrows, nets and spear-throwers), thereby minimizing the need to run long distances. But Carrier⁹ has hypothesized that ER evolved in early hominids for predator pursuit before these inventions in the Upper Palaeolithic (about 40,000 yr ago). ER may have helped hunters get close enough to throw projectiles, or perhaps even to run some mammals to exhaustion in the heat. Although such demanding strategies have been occasionally documented among modern foragers (see ref. 61), they might have been too energetically expensive and low-yield for the benefits to have outweighed the costs.

Another hypothesis to explore is that ER was initially useful for effective scavenging in the open, semi-arid environments apparently inhabited by early *Homo*. If early hominids were regularly scavenging marrow, brain and other tissues from carcasses, then ER would have helped hominids to compete more effectively for these scattered and ephemeral resources. Wild dogs and hyenas often rely upon remote olfactory or visual cues such as circling vultures to identify scavenging opportunities, and then run long distances to secure them^{13,14}. Early *Homo* may thus have needed to run long distances to compete with other scavengers, including other hominids. This hypothesis is difficult to test because modern hunter-gatherers tend to scavenge only opportunistically. However, similar strategies of 'pirating' meat from carnivores are sometimes practised by the Hadza in East Africa⁶² and perhaps were more common in open habitats before the invention of technologies such as the bow and arrow.

Additional research will help to clarify and test when and how ER capabilities evolved in humans, and to examine more thoroughly their implications for human evolution. For example, it is known that major increases in encephalization occurred only after the appearance of early *Homo*^{4,8}. The hypothesis that ER evolved in *Homo* for scavenging or even hunting therefore suggests that ER may have made possible a diet rich in fats and proteins thought to account for the unique human combination of large bodies, small guts, big brains and small teeth^{52,63}. Today, ER is primarily a form of exercise and recreation, but its roots may be as ancient as the origin of the human genus, and its demands a major contributing factor to the human body form. □

doi:10.1038/nature03052.

- Haile-Selassie, Y. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–181 (2001).
- Galik, Y. *et al.* External and internal morphology of the BAR 1002/00 *Orrorin tugenensis* femur. *Science* **305**, 1450–1453 (2004).
- Ward, C. V. Interpreting the posture and locomotion of *Australopithecus afarensis*: where do we stand? *Yb. Physical Anthropol.* **35**, 185–215 (2002).
- Aiello, L. & Dean, M. C. *An Introduction to Human Evolutionary Anatomy* (Academic, London, 1990).
- Rose, M. D. in *Origine(s) de la Bipedie chez les Hominides* (eds Coppens, Y. & Senut, B.) 37–49 (CNRS, Paris, 1991).
- Jungers, W. L. Relative joint size and hominid locomotor adaptations with implications for the evolution of hominid bipedalism. *J. Hum. Evol.* **17**, 247–265 (1988).
- Ruff, C. B. *et al.* in *Primate Locomotion: Recent Advances* (ed. Strasser, E.) 449–469 (Plenum, New York, 1998).
- Wood, B. & Collard, M. The human genus. *Science* **284**, 65–71 (1999).
- Carrier, D. R. The energetic paradox of human running and hominid evolution. *Curr. Anthropol.* **25**, 483–495 (1984).
- Heinrich, B. *Why We Run: A Natural History* (Harper Collins, New York, 2002).
- Garland, T. Jr. The relation between maximal running speed and body-mass in terrestrial mammals. *J. Zool.* **199**, 157–170 (1983).
- Taylor, C. R., Heglund, N. C. & Maloiy, G. M. Energetics and mechanics of terrestrial locomotion. I. Metabolic energy consumption as a function of speed and body size in birds and mammals. *J. Exp. Biol.* **97**, 1–21 (1982).
- Pennycuik, C. J. in *Serengeti: Dynamics of an Ecosystem* (eds Sinclair, A. R. E. & Norton-Griffiths, M.) 164–184 (Univ. Chicago Press, Chicago, 1979).

14. Holekamp, K. E., Boydston, E. E. & Smale, E. in *How and Why Animals Travel in Groups* (eds Boinski, S. & Garber, P.) 587–627 (Univ. Chicago Press, Chicago, 2000).
15. Alexander, R. M. Optimum walking techniques for quadrupeds and bipeds. *J. Zool. Lond.* **192**, 97–117 (1980).
16. Margaria, R., Cerretelli, P., Aghemo, P. & Sassi, G. Energy cost of running. *J. Appl. Physiol.* **18**, 367–370 (1963).
17. Alexander, R. M. Energy-saving mechanisms in walking and running. *J. Exp. Biol.* **160**, 55–69 (1991).
18. Cavagna, G. A., Thys, H. & Zamboni, A. The sources of external work in level walking and running. *J. Physiol. Lond.* **262**, 639–657 (1976).
19. Ker, R. F. *et al.* The spring in the arch of the human foot. *Nature* **325**, 147–149 (1987).
20. Farley, C. T., Glasheen, J. & McMahon, T. A. Running and springs: speed and animal size. *J. Exp. Biol.* **185**, 71–86 (1993).
21. Cavanagh, P. R. & Kram, R. Stride length in distance running: velocity, body dimensions, and added mass effects. *Med. Sci. Sports Exerc.* **21**, 467–479 (1989).
22. Hunt, K. D. Mechanical implications of chimpanzee positional behavior. *Am. J. Phys. Anthropol.* **86**, 521–536 (1991).
23. Isbell, L. A. *et al.* Locomotor activity differences between sympatric patas monkeys (*Erythrocebus patas*) and vervet monkeys (*Cercopithecus aethiops*): implications for the evolution of long hindlimb length in *Homo*. *Am. J. Phys. Anthropol.* **105**, 199–207 (1998).
24. Alexander, R. M., Jayes, A. S. & Ker, R. F. Estimates of energy cost for quadrupedal running gaits. *J. Zool. Lond.* **190**, 155–192 (1980).
25. Heglund, N. C. & Taylor, C. R. Speed, stride frequency and energy cost per stride. How do they change with body size and gait? *J. Exp. Biol.* **138**, 301–318 (1988).
26. Hoyt, D. F. & Taylor, C. R. Gait and the energetics of locomotion in horses. *Nature* **292**, 239–240 (1981).
27. Minetti, A. E. Physiology: efficiency of equine express postal systems. *Nature* **426**, 785–786 (2003).
28. National Center for Chronic Disease Prevention and Health Promotion. *Participation in Physical Activities: Adults Aged 18 and Over* (National Health Information Survey, 1998, now age-adjusted to 2000 population); (<http://www.cdc.gov/nccdphp/dnpa/physical/stats/pasports.htm>).
29. Taylor, C. R. & Rowntree, V. J. Running on two or on four legs: which consumes more energy? *Science* **179**, 186–187 (1973).
30. Thorpe, S. K. *et al.* Dimensions and moment arms of the hind- and forelimb muscles of common chimpanzees (*Pan troglodytes*). *Am. J. Phys. Anthropol.* **110**, 179–199 (1999).
31. Swindler, D. R. & Wood, C. D. *An Atlas of Primate Gross Anatomy: Baboon, Chimpanzee and Man* (Univ. Washington Press, Seattle, 1973).
32. Susman, R. L., Stern, J. T. & Jungers, W. L. Arboreality and bipedality in the Hadar hominids. *Folia Primatol.* **43**, 113–156 (1984).
33. Latimer, B. & Lovejoy, C. O. The calcaneus of *Australopithecus afarensis* and its implications for the evolution of bipedality. *Am. J. Phys. Anthropol.* **78**, 369–386 (1989).
34. Stern, J. T. & Susman, R. L. The locomotor anatomy of *Australopithecus afarensis*. *Am. J. Phys. Anthropol.* **60**, 279–317 (1983).
35. Clarke, R. J. & Tobias, P. V. Sterkfontein Member 2 foot bones of the oldest South African hominid. *Science* **269**, 521–524 (1995).
36. Harcourt-Smith, W. E. H. *Form and Function in the Hominoid Tarsal Skeleton* Thesis, Univ. College London (2002).
37. Lewis, O. J. *Functional Morphology of the Evolving Hand and Foot* (Oxford Univ. Press, Oxford, 1989).
38. Reynolds, T. R. Stride length and its determinants in humans, early hominids, primates, and mammals. *Am. J. Phys. Anthropol.* **72**, 101–115 (1987).
39. Kram, R. & Taylor, C. R. Energetics of running: a new perspective. *Nature* **346**, 265–267 (1990).
40. Haeusler, M. & McHenry, H. M. Body proportions of *Homo habilis* reviewed. *J. Hum. Evol.* **46**, 433–465 (2004).
41. Myers, M. J. & Steudel, K. Effect of limb mass and its distribution on the energetic cost of running. *J. Exp. Biol.* **116**, 363–373 (1985).
42. Zihlman, A. L. & Bruner, L. Hominid bipedalism: then and now. *Yb. Physical Anthropol.* **22**, 132–162 (1979).
43. Mero, A., Jaakkola, L. & Komi, P. V. Relationships between muscle fibre characteristics and physical performance capacity in trained athletic boys. *J. Sports Sci.* **9**, 161–171 (1991).
44. Yang, N. *et al.* ACTN3 genotype is associated with human elite athletic performance. *Am. J. Hum. Genet.* **73**, 627–631 (2003).
45. Keller, T. S. *et al.* Relationship between vertical ground reaction force and speed during walking, slow jogging, and running. *Clin. Biomech.* **11**, 253–259 (1996).
46. Rose, M. D. A hominine hip bone, KNM-ER 3228, from East Lake Turkana, Kenya. *Am. J. Phys. Anthropol.* **63**, 371–378 (1984).
47. Sanders, W. J. Comparative morphometric study of the australopithecine vertebral series Stw-H8/H41. *J. Hum. Evol.* **34**, 249–302 (1998).
48. Biewener, A. A. & Taylor, C. R. Bone strain: a determinant of gait and speed? *J. Exp. Biol.* **123**, 383–400 (1986).
49. Hinrichs, R. N. in *Biomechanics of Distance Running* (ed. Cavanagh, P. R.) 107–133 (Human Kinetics Books, Champaign, Illinois, 1990).
50. McLay, I. S., Lake, M. J. & Cavanagh, P. R. in *Biomechanics of Distance Running* (ed. Cavanagh, P. R.) 165–186 (Human Kinetics Books, Champaign, Illinois, 1990).
51. Jellema, L. M., Latimer, B. & Walker, A. in *The Nariokotome Homo erectus Skeleton* (eds Walker, A. & Leakey, R. E. F.) 294–325 (Harvard Univ. Press, Cambridge, 1993).
52. Aiello, L. & Wheeler, P. The expensive tissue hypothesis: the brain and digestive system in human and primate evolution. *Curr. Anthropol.* **36**, 199–221 (1995).
53. Spoor, F., Wood, B. & Zonneveld, F. Implications of early hominid labyrinthine morphology for evolution of human bipedal locomotion. *Nature* **369**, 645–648 (1994).
54. Bianchi, M. The thickness, shape and arrangement of the elastic fibres within the nuchal ligament from various animal species. *Anat. Anz. Jena* **169**, 53–66 (1989).
55. Wheeler, P. E. The thermoregulatory advantages of hominid bipedalism in open equatorial environments: the contribution of increased convective heat loss and cutaneous evaporative cooling. *J. Hum. Evol.* **21**, 107–115 (1991).
56. Ruff, C. B. Climate and body shape in hominid evolution. *J. Hum. Evol.* **21**, 81–105 (1990).
57. Falk, D. Brain evolution of *Homo*: the radiator theory. *Behav. Brain Sci.* **13**, 333–381 (1990).
58. Cabanac, M. & Caputa, M. Natural selective cooling of the human brain: evidence of its occurrence and magnitude. *J. Physiol. Lond.* **286**, 255–264 (1979).
59. Niinimaa, V., Cole, P., Mintz, S. & Shephard, R. J. The switching point from nasal to oronasal breathing. *Resp. Physiol.* **42**, 61–71 (1981).
60. Semaw, S. *et al.* 2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. *J. Hum. Evol.* **45**, 169–177 (2003).
61. Nabokov, P. *Indian Running: Native American History and Tradition* (Ancient City, Santa Fe, New Mexico, 1987).
62. O'Connell, J. F., Hawkes, K. & Blurton-Jones, N. G. Hadza scavenging: implications for Plio-Pleistocene hominid subsistence. *Curr. Anthropol.* **29**, 356–363 (1988).
63. Wrangham, R. W. *et al.* The raw and the stolen: cooking and the ecology of human origins. *Curr. Anthropol.* **5**, 567–594 (1999).
64. Muybridge, E. *The Human Figure in Motion* (Dover, New York, 1985).
65. Walker, A. & Leakey, R. E. F. *The Nariokotome Homo erectus skeleton* (Harvard Univ. Press, Cambridge, 1993).
66. Lovejoy, O. Evolution of human walking. *Sci. Am.* **259**, 118–125 (1988).
67. Dillman, C. J. Kinematic analyses of running. *Exercise Sports Sci. Rev.* **3**, 193–218 (1975).

Acknowledgements We are grateful to A. Biewener, D. Carrier, W. Harcourt-Smith, F. Jenkins, Jr, J. McGrath, D. Pilbeam, J. Polk, H. Pontzer and R. Wrangham for discussion and comments on the manuscript. Funding was provided by the American School of Prehistoric Research; illustrations in Fig. 4 were rendered by L. Meszoly. D.M.B. and D.E.L. contributed equally to this work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.M.B. (bramble@bioscience.utah.edu) or D.E.L. (danlieb@fas.harvard.edu).

Regulation of p53 activity through lysine methylation

Sergei Chuikov¹, Julia K. Kurash², Jonathan R. Wilson³, Bing Xiao³, Neil Justin³, Gleb S. Ivanov², Kristine McKinney⁴, Paul Tempst⁵, Carol Prives⁴, Steven J. Gambelin³, Nikolai A. Barlev² & Danny Reinberg¹

¹Howard Hughes Medical Institute, Division of Nucleic Acids Enzymology, Department of Biochemistry, Robert Wood Johnson Medical School, Piscataway, New Jersey 08854, USA

²Molecular Oncology Research Institute, NEMC-Tufts School of Medicine, 75 Kneeland Street, Boston, Massachusetts 02111, USA

³Structural Biology Group, MRC-NIMR, The Ridgeway, Mill Hill, London NW7 1AA, UK

⁴Department of Biological Sciences, Columbia University, New York, New York 10027, USA

⁵Molecular Biology Program, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

p53 is a tumour suppressor that regulates the cellular response to genotoxic stresses. p53 is a short-lived protein and its activity is regulated mostly by stabilization via different post-translational modifications. Here we report a novel mechanism of p53 regulation through lysine methylation by Set9 methyltransferase. Set9 specifically methylates p53 at one residue within the carboxyl-terminus regulatory region. Methylated p53 is restricted to the nucleus and the modification positively affects its stability. Set9 regulates the expression of p53 target genes in a manner dependent on the p53-methylation site. The crystal structure of a ternary complex of Set9 with a p53 peptide and the cofactor product S-adenosyl-L-homocysteine (AdoHcy) provides the molecular basis for recognition of p53 by this lysine methyltransferase.

Various histone modifications such as phosphorylation, acetylation, methylation and ubiquitination have been implicated in the regulation of gene expression by several mechanisms. The interplay between these modifications has been studied in detail and a 'histone code' hypothesis has been proposed to explain the effects of these modifications on gene expression¹. Recently, we identified a histone methyltransferase, Set9, that targets histone H3 at lysine 4 and showed that this modification is associated with gene activation^{2,3}. Here we demonstrate that Set9 activity is not limited to histones, it is also able to methylate the tumour suppressor p53.

p53 is a transcription factor that is mutated in approximately 50% of human cancers. In normal cells, p53 exerts a pivotal role in controlling the cell cycle, apoptosis and DNA repair in response to various forms of genotoxic stress. The regulation of p53 is complex and occurs mainly at the post-translational level⁴. This complexity is realized through the number and types of different post-translational modifications that contribute to its stabilization and activation. Phosphorylation of several residues at the amino terminus of p53 has been shown to affect its interaction with MDM2, which targets p53 for ubiquitin-mediated degradation^{5–7}. The C terminus of p53 is rich in lysines, which are subjected to acetylation, ubiquitination and sumoylation⁴. Acetylation of the C terminus has been shown to protect p53 from ubiquitination⁸. Moreover, acetylation of p53 at lysines 373 and 382 increases its DNA-binding activity^{9,10} and potentiates its interaction with other transcription factors¹¹. The positive effects of acetylation on p53 activity can be reversed by deacetylation¹². p53 has also been shown to be sumoylated at K386, although the exact role of this modification in the regulation of p53 is not yet clear^{13–15}. Taken together, these data suggest very complex regulatory mechanisms, the coordination of which still remains elusive.

p53 methylation by a lysine-specific methyltransferase Set9

Several enzymes that methylate specific lysine residues within the histone tails—histone lysine methyltransferases (HKMTs)—have been isolated and characterized. One such enzyme is Set9 (also known as Set7), which targets lysine 4 of histone H3 (H3-K4) *in vitro*^{2,3}. However, purified Set9 was unable to methylate H3-K4 assembled into nucleosomes (Fig. 1a), which are thought to be the

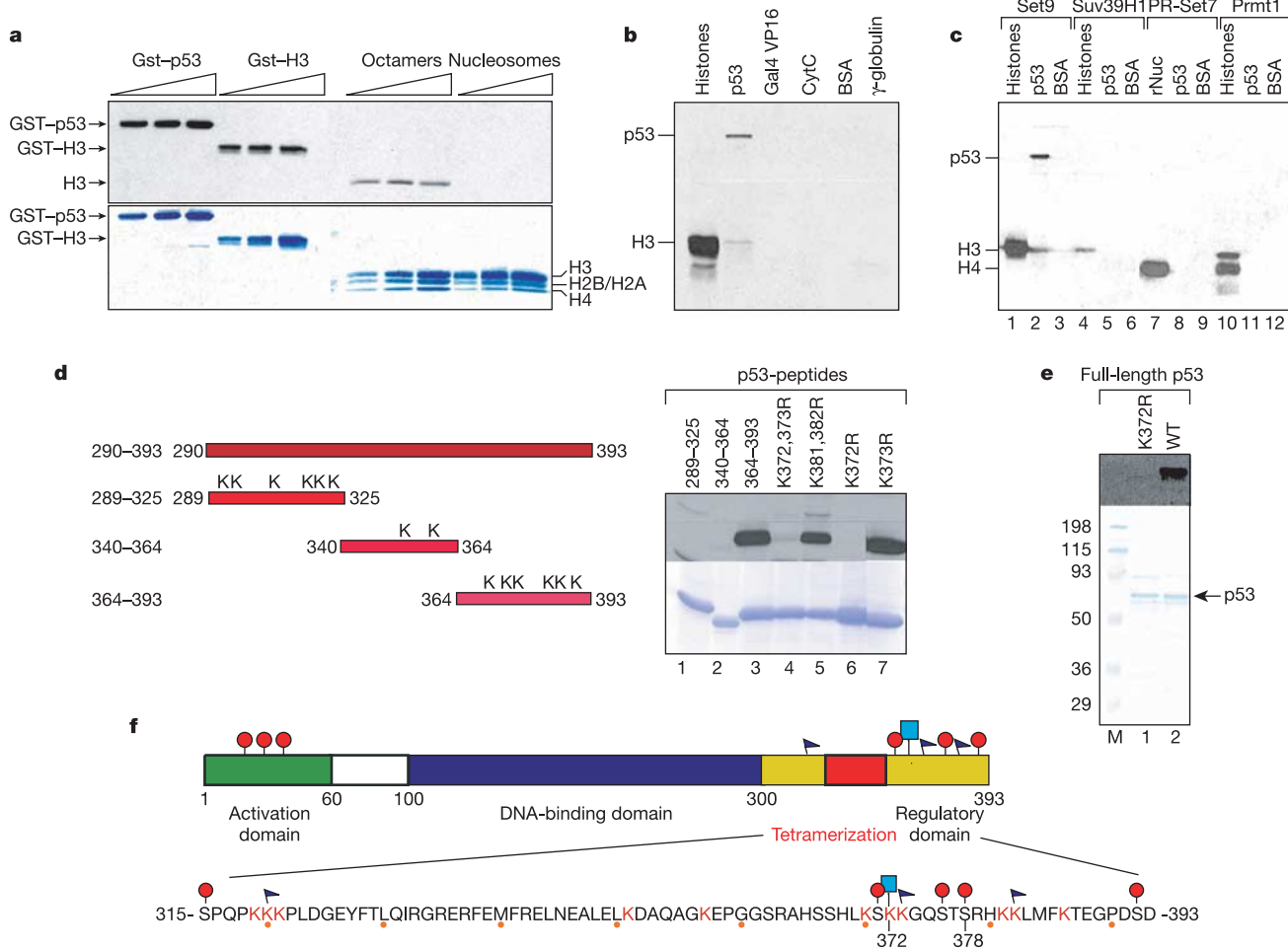
physiologically relevant substrates for chromatin-modifying enzymes. However, methylation of nucleosomal H3-K4 does exist *in vivo*, but is catalysed by other enzymes, such as Set1 in yeast and by the related Set1, Ash1, SMYD3, trithorax and MLL proteins in higher eukaryotes. We therefore searched for additional candidate protein substrates, other than the histone polypeptides, that may be targeted by Set9 for methylation. We found that, in addition to several polypeptides (data not shown), Set9 methylated the tumour suppressor protein p53 *in vitro* (Fig. 1a). Under our assay conditions Set9-mediated methylation of p53 appeared to be quite specific. Other polypeptides such as cytochrome c, Gal4-VP16, bovine serum albumin (BSA) and γ -globulin did not serve as Set9 substrates (Fig. 1b). Moreover, other protein methyltransferases, such as Suv39H1, which targets H3-K9, PR-Set7 that targets H4-K20, or the arginine methyltransferase PRMT1, failed to use p53 as a substrate (Fig. 1c). Interestingly, recent studies have indicated that Set9 can methylate TAF10 (ref. 16), and although our studies did not address TAF10, the results collectively suggest that Set9-mediated methylation of proteins other than histone may be more general.

To gain insight into the role of p53 methylation, we first mapped the residue of p53 methylated by Set9. The initial analysis was performed using three fragments of p53 encompassing its functional domains: the N terminus (residues 1–82) representing the transactivation domain; the middle part of the protein containing the DNA-binding domain (residues 96–312); and the regulatory C terminus of the protein (residues 290–393). Set9 methylated p53 within the regulatory C terminus (data not shown). Next, the C terminus was further divided into three protein fragments (Fig. 1d, left panel) and each was analysed for methylation by Set9. The most C-terminal fragment, extending over amino acids 363 to 394, was a substrate for methylation (Fig. 1d, right panel, compare lanes 1–3). This polypeptide contained six lysines, which were then substituted individually, or in combination, with non-methylatable arginines and tested as substrates for methylation. We found that Set9 methylated lysine 372 (lanes 4–7). Consistent with this observation, a single substitution at lysine 372 with arginine in the full-length p53 protein eliminated methylation by Set9 (Fig. 1e). These data established that the site of Set9 methylation is within a region of the

[illegible]

TABLE 1. *Estimated values of the parameters of the model for the 1997-1998 season*

The overall structure of the complex, shown in ribbon represen-



d. Schematic representation of the C-terminal p52 peptides and lysine candidates for

ternary structures with p53 or H3 substrates shows a strong overall similarity between the two (ribbons representation in Supplementary Fig. S2). Importantly, the mono-methylated lysine side-chain of the p53 peptide accesses the methyl donor cofactor using the same narrow channel running through the SET domain as that described for the H3 complex. Overall, the electron-density map for the Set9–p53 complex is of high quality. There is well-resolved electron density for the first six residues of the p53 peptide, for residues Leu (–3) to Gly (+2) (numbering is with respect to the modified p53 lysine residue 372) but the last four residues of the peptide are disordered. This observation suggests that Set9 interacts with a rather short recognition sequence at the active site of the enzyme. The Set9 residues involved in interactions with the three amino acids that are N-terminal to the modified lysine are the same for the p53 peptide and the H3 peptide despite the difference in peptide sequence (Fig. 3d). The hydrogen bond between the carbonyl of Arg 258 and the Ne atom of Arg (–2) in the H3 complex is replaced by a hydrogen bond with the Ne of Lys (–2) in p53. The hydrophobic packing of Trp 260 with the peptide Arg (–2) in H3 is substituted with the packing against Lys (–2) in p53. There is also a conservative substitution of Ser (–1) in H3 for Thr (–1) in p53 that maintains the hydrogen bond with Ser 268 of the enzyme.

Of the five Set9 residues making polar contacts with the peptide, four of them—Asp 256, Arg 258, Thr 266 and Ser 268—are located

within the variable Set-I region. The sixth residue making a polar contact with the peptide is located in the C-flanking domain (Tyr 335). Together, these observations highlight the significance of the variable Set-I and C-flanking domains in determining the specificity of the SET enzymes with respect to which lysine residue within a stretch of polybasic residues will be modified. The binding studies reported before also indicated that there must be amino-acid residues located at some distance from the target lysine residue that play an important role in determining substrate specificity. Although no structural data are available for the interactions formed between these more distant residues with Set9, it is tempting to speculate that the N-flanking domain, preceding the SET domain, may be involved in mediating these interactions.

p53 is methylated *in vivo*

To facilitate studies on the role of Set9-methylation of p53 *in vivo*, we generated a polyclonal antibody that specifically recognized mono-methylated p53-K372. Rabbit antibodies were initially screened against p53 peptides harbouring K372 with different degrees of methylation. An antibody that specifically recognized a mono-methylated peptide, but failed to detect an unmodified peptide or equivalent peptides that were di- or tri-methylated was selected (Fig. 3a). The anti-p53-K372-mono-methyl (anti-p53-K372me) antibody specifically detected bacterially purified p53

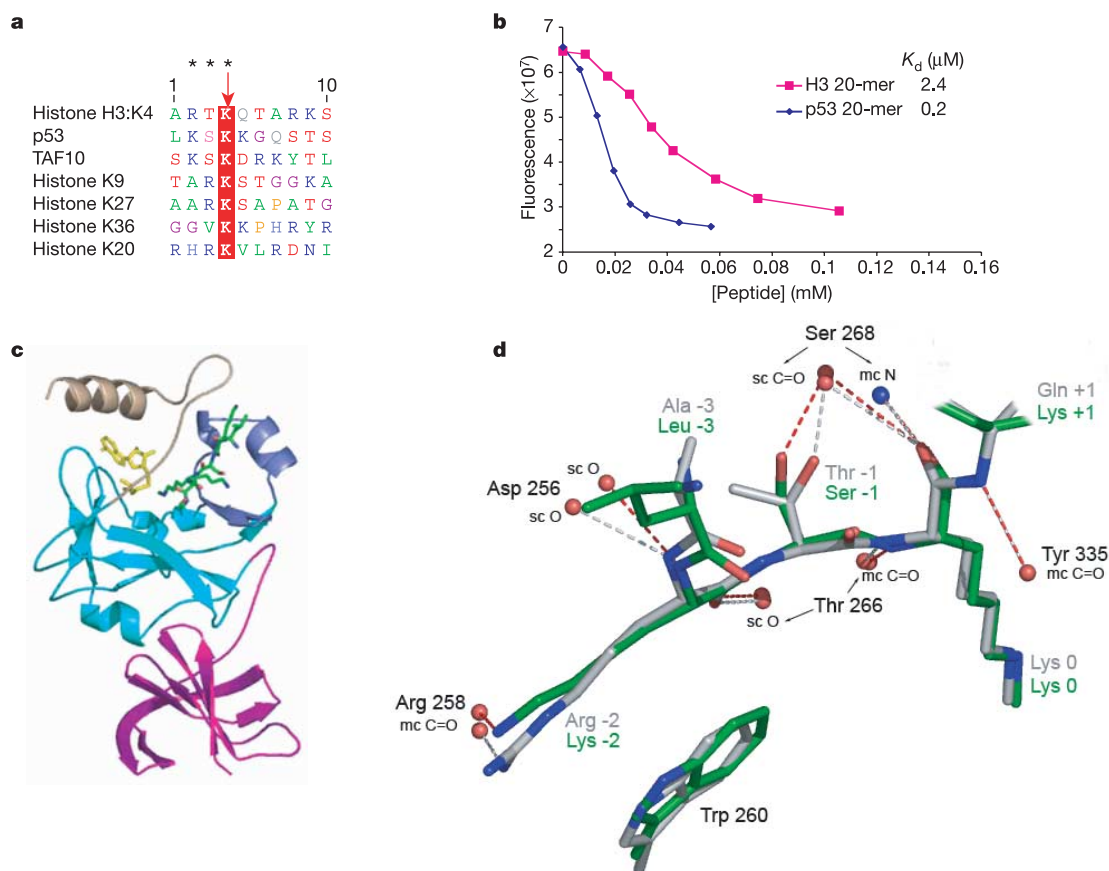


Figure 2 Interaction of Set9 with p53 and H3 peptides. **a**, Alignment of protein sequences adjacent to lysines targeted for methylation in the case of histone H3 lysine 4, p53, TAF10, histone H3 lysine 9, histone H3 lysine 27, histone H3 lysine 36, and histone H4 lysine 20. Methylated lysine is highlighted in red and the asterisk represents the consensus for substrate recognition by Set9 methyltransferase. **b**, The dissociation constants of the p53 and H3 peptides, shown in inset, were determined using fluorometric competition assay. The unlabelled p53 and H3 peptides were used to displace a dansyl-labelled H3 10-mer of known affinity. The displacement curves were used to calculate the dissociation

constants of the unlabelled peptides (shown in inset). **c**, Overall structure of the Set9–p53 complex. Magenta, N-flanking domain; cyan, Set domain; blue, Set-I domain; beige, C-flanking domain. The backbone of the p53 peptide is green, and AdoHcy is yellow. **d**, Overlay of the peptide showing enzyme interactions for the Set9–p53 complex and the Set9/H3 complex. The p53 carbon atoms are shown in green and H3 in grey. Hydrogen bonds are represented by dashed lines: only the Set9 donor/acceptor atoms are shown for each complex.

protein methylated by Set9 *in vitro*, but was severely compromised with the untreated p53 protein. Equivalent amounts of methylated and mock-methylated p53 were used in the western analysis (Fig. 3b). We concluded that the antibody specifically recognized K372-methylated p53.

To address the question of whether p53 is methylated by Set9 *in vivo*, wild-type Set9 or its catalytically inactive mutant form (H297A) were stably transfected into 293F cells expressing endogenous p53. Western-blot analysis using anti-p53-K372me antibody detected methylated p53 protein in extracts derived from cells overexpressing the wild-type Set9 protein, but cells overexpressing the catalytically inactive Set9 protein displayed little methylated p53 (Fig. 3c). The expression levels of Set9 proteins and the amounts of p53 protein in the extracts were similar (Fig. 3c, bottom panels). We therefore concluded that transfected Set9 methylated p53 in 293F cells.

We next examined p53 methylation at lysine 372 *in vivo* under conditions in which the intracellular concentration of Set9 was unaltered. 293F cells were treated with adriamycin (Adr), a chemical that induces DNA damage resulting in the activation of a p53-responsive pathway. Cell extracts prepared from untreated and Adr-treated cells, were then compared for total amounts of p53 protein using monoclonal p53 antibody. Under these conditions, DNA damage did not increase the levels of intracellular p53; this is not surprising because p53 is already stabilized by adenoviral proteins expressed endogenously in 293F cells¹⁸. Both extracts were also subjected to immunoprecipitation with anti-p53-K372me antibodies followed by western blotting with p53-specific antibody. Adr treatment increased the intracellular concentration of methylated K372-p53 protein (Fig. 3d). Importantly, the levels of Set9 remained unchanged after Adr treatment (data not shown). p53 also became methylated in response to DNA damage in U2OS cells (Fig. 3e). Importantly, the introduction of Set9 short interfering (si) RNA in these cells decreased K372 methylation of p53 upon

Adr treatment (Fig. 3e) and correlated with the decreased levels of Set9 (see Fig. 4c). These findings collectively suggested that methylation of p53 in response to DNA damage is a general phenomenon, rather than a cell-type-specific event.

Methylated p53 is in the nucleus and regulates p53 target genes

In our initial attempt to characterize the function of p53-K372 methylation, we stably transfected 293F cells with a Set9 expression vector. Extracts from these cells were then fractionated into nuclear and cytosolic fractions to determine the distribution of methylated p53-K372. These studies suggested that all the methylated p53-K372 was localized to the nucleus, although p53 was equally distributed between the nuclear and cytosolic fractions (Fig. 4a). Importantly, untransfected 293F cells also displayed methylated p53 in the nuclear fraction, ruling out the possibility that the nuclear localization of methylated p53-K372 is a consequence of Set9 overexpression.

We next analysed whether Set9-mediated methylation of p53 affected its transcriptional activity. We focused on one of the best-characterized p53 transcriptional targets, the *p21/WAF/CIP* gene, whose product is critical for cell-cycle arrest in the G1 phase. The effect of Set9-mediated p53-K372 methylation on transcription of the endogenous p21 gene was measured in U2OS cells using reverse-transcription polymerase chain reaction (RT-PCR). Overexpression of Set9 (Fig. 4b, middle panel), resulted in increased expression of p21 (Fig. 4b, left panel), and correlated with increased levels of K372-methylated p53 (Fig. 4b, right panel). Importantly, even in the absence of DNA damage, the levels of p21 expression in cells overproducing Set9 were higher than those in the parental non-transfected cells treated with the DNA-damaging agent (Fig. 4b, compare lane 4 with lane 1). As expected, overexpression of Set9 followed by Adr treatment resulted in a further increase in the levels of p21 expression (Fig. 4b, lane 3).

We directly analysed the contribution of Set9 to regulation of p21

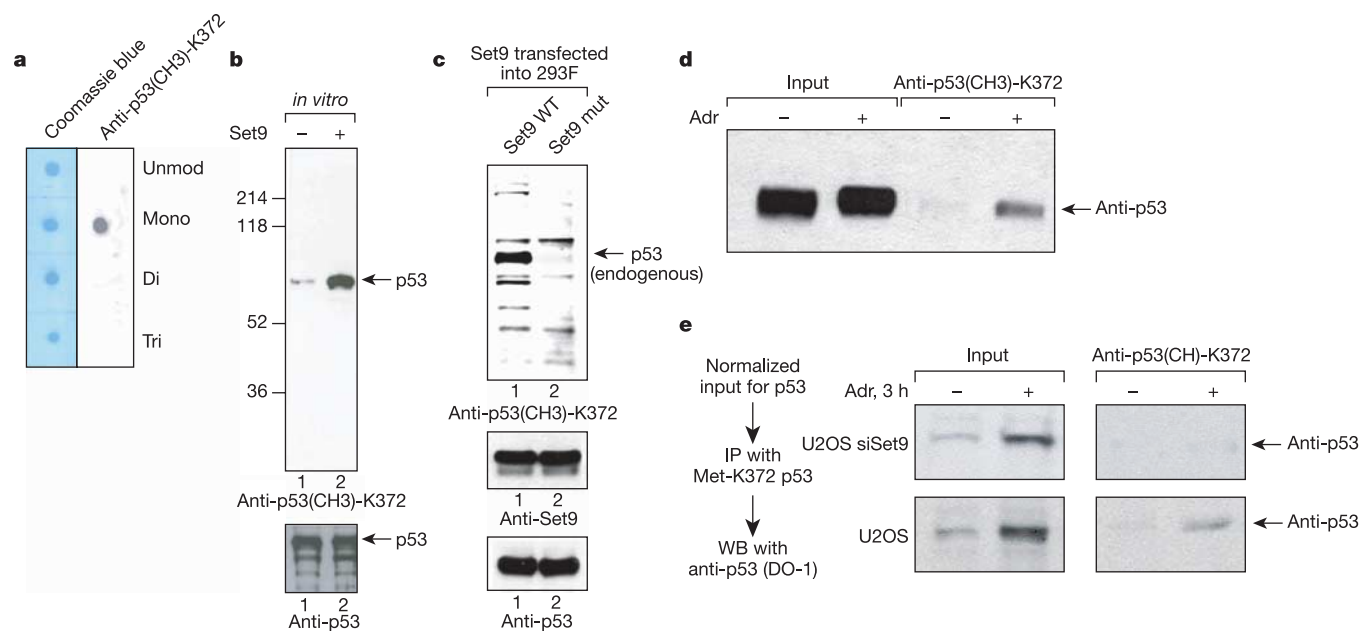


Figure 3 Set9 methylates p53 *in vivo*. **a**, Dot blot assay of the p53 peptides containing various degrees of methylation. Equal amounts of unmodified, mono-, di- and tri-methylated p53 peptides were immunoblotted with mono-methyl p53 antibody (right panel). Coomassie-blue staining of these peptides served as a loading control (left). **b**, Western-blot analysis of recombinant p53 methylated by Set9 *in vitro* using mono-methyl p53 antibody (top). Western-blot signal using anti-p53 antibody served as a loading control (bottom). **c**, Detection of endogenous methylated p53 from whole-cell extracts of 293F stable cell lines overexpressing wild-type (WT) or mutant (mut) Set9.

Methylated p53 was detected by western blotting with anti-p53K372me antibody (top). Western-blot signals obtained with anti-Set9 and anti-p53 antibodies served as loading controls. **d**, Immunoprecipitation of methylated endogenous p53 from nuclear extracts of 293F cells treated or mock-treated with Adr with anti-p53K372me Ab. Western blot with anti-p53 antibody (DO-1) is also shown. **e**, U2OS cells with or without Set9 siRNA were treated with 0.5 μ M of Adr for 3 h. Nuclear extracts from treated or mock-treated cells were normalized for the amounts of total p53 and subjected to immunoprecipitation (IP) with anti-p53K372me antibody. Western blot with anti-p53 antibody (DO-1) is shown.

expression. The levels of p21 expression in U2OS cells stably expressing ectopic Set9 were approximately twofold higher than in the parental strain. Consistent with the results presented in Fig. 4b, a similar pattern of p21 expression was observed upon treatment of cells with 5FU (a compound that also induces DNA damage), that is, in the presence of Set9 overexpression the levels of p21 increased over twofold compared to the parental cells. Strikingly, overexpression of a catalytically inactive form of Set9 completely ablated the induction of p21 in response to DNA damage (Fig. 4c). Most importantly, si-RNA-mediated reduction of the intracellular levels of Set9 also impaired p21 expression upon DNA damage in U2OS cells (Fig. 4c). These findings collectively established that Set9 is required, directly or indirectly, for the expression of p21. Surprisingly, the introduction of either the Set9 mutant or si-RNA against Set9 protein also decreased the total levels of p53, without affecting

the expression of unrelated proteins, such as tubulin (Figs 4c and 5b). The observed stimulatory effect of Set9 on p21 gene expression extended to other p53-responsive genes such as BAX and MDM2 (Supplementary Fig. S3).

Increased stability of methylated p53-K372

That the intracellular levels of p53 were critically dependent on the enzymatic activity of Set9 suggested that methylation of p53-K372 might affect p53 stability. We directly tested this hypothesis with pulse-chase experiments. U2OS cells were transiently transfected with wild-type or a catalytically inactive Set9 and grown in a methionine-free medium in the presence of ^{35}S -methionine for 30 min. The cells were then washed, resuspended in media containing methionine and then a fraction of the cells were withdrawn at the indicated times. The amount of ^{35}S -radiolabelled p53 in the

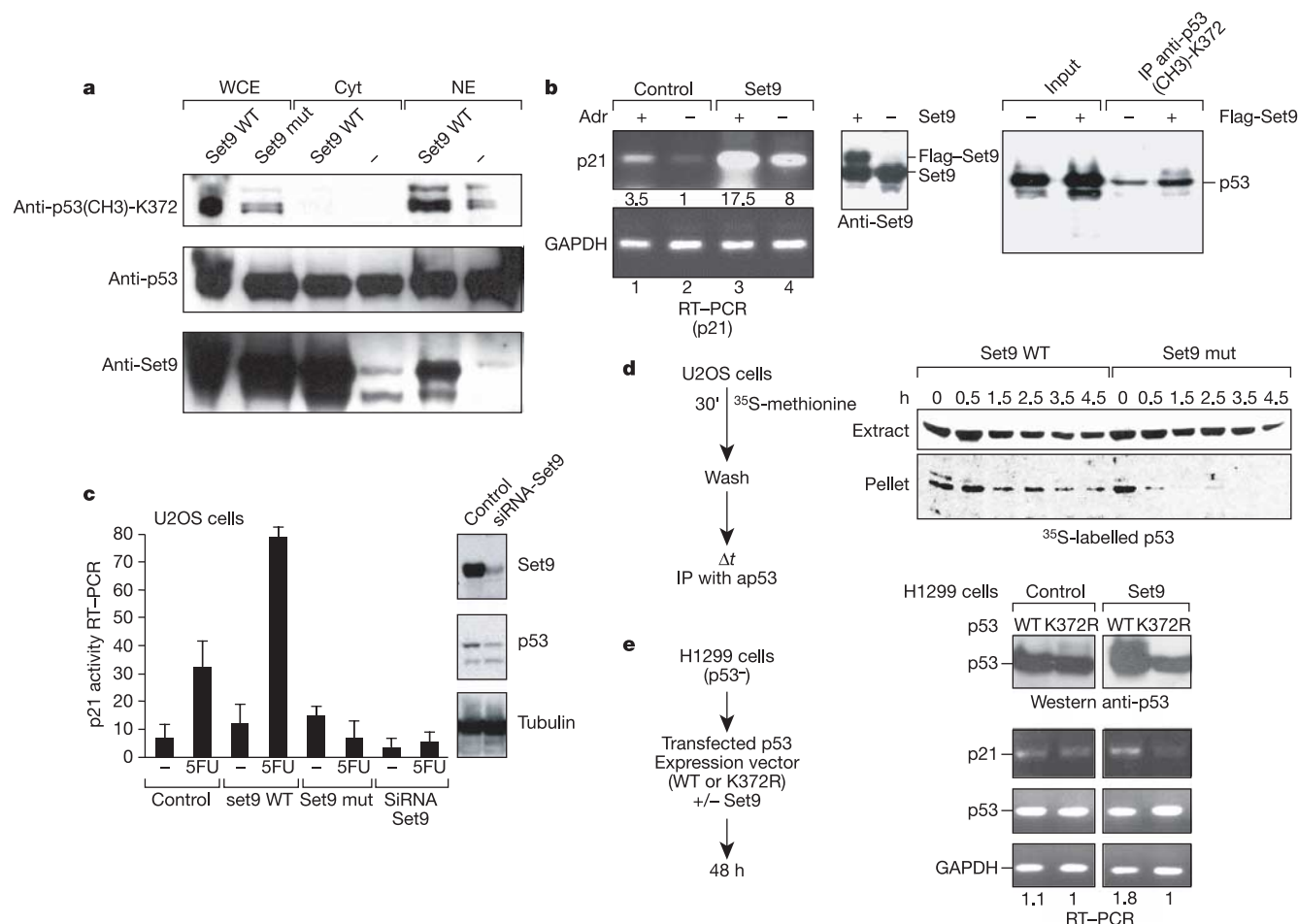


Figure 4 Set9 potentiates p53 function. **a**, 293F cells stably expressing Set9 were lysed and whole-cell extract, cytoplasmic and nuclear fractions were prepared. Top, western-blot analysis of methylated p53 with anti-p53K372me antibody; middle, western-blot analysis of total p53 with anti-p53 antibody (served as a loading control); bottom, western-blot analysis of Set9. **b**, RT-PCR of p21 gene expression in U2OS cells with normal or overexpressed amounts of Set9 after treatment with the DNA-damaging agent Adr (left). Shown are the relative values normalized to the p21 signal in untreated U2OS cell. U2OS cells were stably transfected with Flag-Set9 wild type, or mock-transfected. Western-blot analysis of the endogenous and ectopic Set9 proteins expressed in U2OS cells. Non-transfected cells were used as control (middle panel). Endogenous p53 was immunoprecipitated with anti-p53K372me antibody from nuclear extracts of U2OS cells transiently transfected with either Flag-Set9 or control constructs. Western-blot analysis of p53 levels in both types of cells with anti-p53 antibody (DO-1) is shown. Input lanes represent 10% of the starting amounts of p53 in both cell lines. IP lanes show the amounts of methylated p53 immunoprecipitated from these cell lines (right panel). **c**, U2OS cells

were stably transfected with Flag-Set9 WT, Flag-Set9 mutant and siRNA against Set9. Non-transfected cells served as control. Cells were mock-treated, or treated with 0.3 μM of 5-fluoro-uracil (5-FU). p21 gene expression was analysed by quantitative real-time PCR using primers specific for the coding region. Shown are the relative values normalized to the GAPDH signal (left). Western-blot analysis of Set9 and p53 in U2OS cells before and after transfection with siRNA against Set9 (right, upper and middle, respectively). Tubulin was used as a loading control (right, bottom). **d**, U2OS cells transiently transfected with Set9 WT or Set9 mutant and labelled with ^{35}S methionine. At the indicated times, cells were lysed and both soluble and insoluble fractions were immunoprecipitated separately with anti-p53 antibody. The immunoprecipitates were resolved by SDS-PAGE and visualized by autoradiography. **e**, Wild-type or K372R mutant of p53 were co-transfected with Set9 WT or control vector into H1299 cells. After 48 h, cells were collected and analysed for p53 protein levels by western blotting (top) and for p53 and p21 gene expression levels by RT-PCR (bottom). The level of GAPDH expression served as control. Shown are the relative values normalized to the levels of p21 activated by p53 K372R.

cytosolic and nuclear pellet fractions was analysed by immunoprecipitation of p53 using monoclonal antibody. The stability of nuclear p53 increased in cells expressing wild-type Set9, but not in cells expressing a catalytically inactive form of Set9 (Fig. 4d). Consistent with our findings that methylated p53-K372 is nuclear (Fig. 3a), p53 stabilization was apparent only in the fraction of nuclear p53 associated with chromatin (Fig. 4d). Thus, we concluded that Set9-mediated methylation of p53-K372 resulted in the stabilization of a chromatin-bound fraction of p53. This conclusion was supported by experiments performed in H1299 cells lacking endogenous p53. Increased stabilization of wild-type p53, but not its methylation-defective mutant (K372R) was observed in these cells upon overexpression of Set9 (Fig. 4e, compare left and right panels). The expression levels of ectopic wild-type and K372R mutant p53 genes were similar, as shown by RT-PCR (Fig. 4e bottom), and thus cannot account for the observed difference in the

p53 protein levels. Moreover, Set9 overexpression resulted in increased expression of the *p21* gene specifically in the presence of wild-type p53, but not with the K372R mutant (Fig. 4e bottom), confirming that Set9 regulates p53-dependent genes through p53 methylation at K372.

We speculated that, given that overexpression of wild-type Set9 resulted in 'hyper-stabilization' and activation of nuclear p53, induction of cell-cycle arrest and apoptosis should ensue. We directly tested this hypothesis by measuring the p53-dependent apoptotic response induced by DNA damage. Consistent with previous reports, treatment of U2OS cells with ADR increased the number of cells positively stained for Annexin V (Fig. 5a, compare panels 1 and 2). Importantly, cells overexpressing Set9 exhibited higher apoptotic staining even without DNA damage (compare panels 3 and 1). The treatment of Set9-overexpressing cells with ADR further increased the Annexin V staining (compare panels 3 and 4). Overexpression of the catalytically inactive Set9 abrogated DNA-damage-induced apoptosis, suggesting that the methyltransferase activity of Set9 is critical for induction of p53-dependent apoptosis (Fig. 5a, compare panels 5 and 6). Again, the overexpression of a catalytically inactive Set9 mutant protein resulted in decreased p53 levels, confirming a functional interaction between p53 and Set9 (Fig. 5b). Importantly, Set9-mediated regulation of cell cycle and apoptosis was p53-dependent, because in the absence of p53 transfected Set9 was unable to induce p21 gene expression in H1299 (Supplementary Fig. S4) or apoptosis in Saos-2 cells in response to DNA damage (Supplementary Fig. S5).

Next, we tested whether p53 methylated by Set9 is present at the promoters of p53 target genes. Our work suggested that p53 methylation is likely to precede acetylation (manuscript in preparation). Therefore, we reasoned that p53 methylation must occur at very early times after DNA damage. To test this hypothesis, we treated normal U2OS and Set9-siRNA-expressing cells with ADR at different times (Fig. 6). At the indicated time periods, cells were harvested, and DNA-protein complexes were cross-linked and subjected to chromatin immunoprecipitation (ChIP) analysis, initially using anti-p53-K372me antibodies. To increase the specificity of ChIP, the anti-p53-K372me immunoprecipitates were eluted from the beads with 1% SDS, diluted with immunoprecipitation buffer and then subjected to a second round of immunoprecipitation using anti-p53 antibody. Consistent with our hypothesis, the amount of methylated p53 bound to the p21 promoter increased as early as 1.5 h after DNA damage. Importantly, K372 methylation was Set9-specific, because U2OS cells expressing siRNA against Set9 did not exhibit a significant increase in methylated p53 at the p21 promoter. These results suggested that methylation of p53, in addition to acetylation and phosphorylation, may represent an important DNA-damage-induced modification mark required for p53 function *in vivo*.

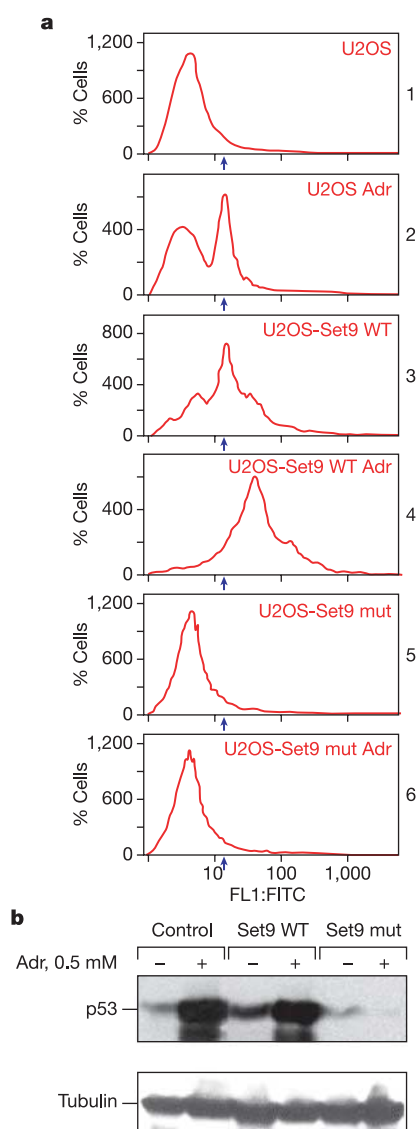


Figure 5 Set9 increases apoptosis in U2OS cells. **a**, The efficiency of apoptosis was determined by Annexin V-fluorescein isothiocyanate (FITC) staining. U2OS stably expressing Set9 WT or Set9 mutant were treated with 0.5 μ M ADR for 24 h or untreated. Non-transfected cells were used as control. Cells were stained with propidium to eliminate necrotic cells and with Annexin V-FITC to score for apoptotic cells. Arrows show the peak of apoptotic cells. **b**, Western-blot analysis of p53 in whole-cell extracts from U2OS cells stably expressing Set9 WT or its catalytically inactive mutant (top). Western-blot signal of tubulin was used as a loading control (bottom).

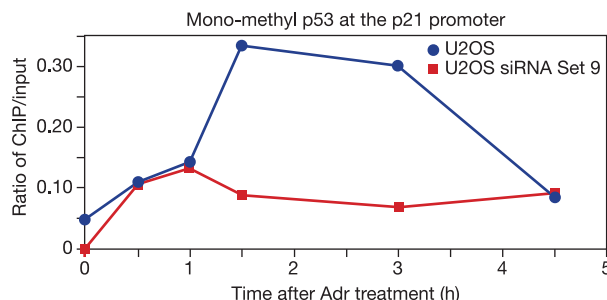


Figure 6 Methylation of p53 at the p21 promoter. U2OS and U2OS Set9 siRNA cells were treated with 0.5 μ M of ADR and collected at the indicated time points for ChIP assay performed using anti-p53K372me antibody. The amounts of precipitated p21 promoter were determined by quantitative RT-PCR as described above.

Here we have established that Set9 methylates p53 at a specific lysine residue *in vivo*. Importantly, we found that this methylation is required for p53 stabilization. What is the mechanism for methylation-induced stabilization of p53? One of the possibilities is that methylation interferes with MDM2-mediated ubiquitination of the six lysine residues at the C terminus, which leads to subsequent p53 degradation and/or nuclear export⁴. Yet, because of its small size, the mono-methylation mark itself is unlikely to affect ubiquitination. We speculate that similar to histones, as-yet-unidentified factors may bind methylated p53 and interfere with Mdm2-dependent ubiquitination. Overexpression of the catalytically inactive Set9 mutant decreased intracellular levels of p53 and attenuated its activity. This confirms the functional connection between Set9 and p53 and also highlights another possible mechanism for p53 inactivation in human cancers. Finally, based on the observation that p53 is a better substrate for Set9 than histone H3, we propose that Set9 might regulate the function of other factors. Likewise, the activities of other known lysine methyltransferases may not be limited to histones, but may also target factors important in many cellular processes. In this regard, the ternary complex of Set9 with p53 reveals the molecular basis for Set9 substrate recognition and may lead to the identification of new Set9 targets in the future. □

Methods

Constructs

GST-H3 N-terminal peptide was expressed as described¹⁹. For GST-C-terminal p53, construct C-terminal (amino acids 290–393) was cloned into PGEX3 vector (Promega). For methylation site identification, different p53 peptides (aa 290–235, 340–364, 364–393) were cloned into Pet102 (Invitrogen) vector and expressed in *Escherichia coli* as a thioredoxin fusion proteins.

Peptides

The following peptides were chemically synthesized for antibody production and dot blots.

p53 unmodified: NH₂-CSHLKSKKGST-COOH; p53 mono-methyl-K372: NH₂-CSHLKSK-(Me)KGST-COOH; p53 di-methyl-K372: NH₂-CSHLKSK-(Me₂)KGST-COOH; p53 tri-methyl-K372: NH₂-CSHLKSK-(Me₃)KGST-COOH.

Methylation assay

Methylation assay was performed as described previously³.

Stability assays

U2OS cells were transfected with Set9 WT or mutant 36 h before labelling. 24 h after transfection, cells were split into 60-mm plates to 50% confluence. Next morning, cells were washed twice with PBS and preincubated for 30 min with DMEM (without methionine) containing 5% dialysed FCS. Cells were labelled for 30 min in fresh methionine-free DMEM containing 0.5 mCi of ³⁵S methionine. Radioactive media were then removed and cells were washed with PBS, followed by the addition of DMEM containing FCS, and 2 mM methionine. At the indicated times cells were washed with PBS and collected. Cytoplasmic and nuclear fractions were prepared as described³ and p53 was immunoprecipitated from each fraction with p53-specific antibody (DO-1).

siRNA construct and Set9 knock-down

For lentiviral vector-based knock-down, the desired 23 base pair (bp) stem-loop RNAs were expressed from the pLSL-GFP vector (a gift from P. Chumakov) driven by the human H1 gene promoter. The vector also contained a minimal histone H4 promoter that drives transcription of a green fluorescent protein (GFP) gene used to monitor infection efficiency. pLSL-GFP vectors expressing the following hairpin RNAs were co-transfected with packaging plasmids pCMV-VSVG and pCMV-deltaR8.2 into 293T cells on a 6-cm dish by standard calcium-phosphate precipitation. Viral supernatants were collected from the transfected 293T cells 24, 36 and 48 h post-transfection, filtered and used for infection of ~5 × 10⁴ target U2OS cells grown on six-well plates in the presence of 4 µg ml⁻¹ of polybrene (Sigma). The U2OS cells were infected with >100% efficiency as judged by GFP fluorescence, expanded and used as a mass culture for all subsequent experiments.

The stem-loop siRNAs were synthesized as two complementary 67-nt oligos, annealed and cloned into *Bam*HI/*Eco*RI sites of pLSL-GFP vector. The siRNA oligo used to target SET9 is:

5' gatccgaccttgagcatgacgattactctctgtcaGTAATCCGTCATCGTCCAGGTGCTctttttg-3' (targeted sequences are in upper case, the loop sequence between sense and antisense 23-mers is in lower case, restriction sites that overhang nucleotides are underlined).

RT-PCR

Total RNA was extracted using the TRI reagent (Invitrogen) according to the manufacturer's instructions. First-strand complementary DNA was synthesized using Ready-To-Go You-Prime First-Strand Beads (Amersham Biosciences).

Real-time PCR

Real-time PCR was performed with the DNA Engine Opticon 2 System (MJ Research) according to the manufacturer's instructions. The primers for human p21 (Cip1/WAF1¹) and GAPDH (GenBank accession numbers NM_000389 and NP_002037 respectively) were designed using the Primer3 program. Primers for the p21 gene: 5'-CACCGAG ACACCACTGGAGG-3' and 5'-GAGAAGATCAGCCGCGCTTT-3', GAPDH: 5'-GGGAAGGTGAAGTCTGGAGT-3' and 5'-TTGAGGTCAATGAAGGGGTCA-3' were synthesized and purified by IDT DNA Technologies. Primer pairs were designed to amplify the appropriate DNA fragments using the following conditions: 10 min initial denaturation, followed by cycles of 1 min at 94 °C, 1 min at 62 °C and 1 min at 72 °C. Fluorescence was measured after the end of the elongation phase at 79 °C. Expression levels were calculated as a ratio between p21 and GAPDH signals. Correct PCR products were confirmed by agarose gel electrophoresis (2% w/v) and melting curve analysis.

ChIP assay

ChIP assay on U2OS and U2OS cells expressing siRNA-Set9 was performed as described previously¹¹. Briefly, cells were cross-linked with 1% formaldehyde, neutralized with 0.125 M glycine and harvested by scraping. The chromatin fraction was prepared by incubating the cells in lysis buffer (10 mM Tris pH 8.0, 200 mM NaCl, 10 mM EDTA, 0.5 mM EGTA, 1 mM PMSF). The insoluble fraction was sonicated and subjected to immunoprecipitation with anti-p53-K372me serum. Following washes, the immunoprecipitated material was eluted with 1% SDS and 10 mM DTT, diluted ten times and re-immunoprecipitated with anti-p53 serum. DNA was then isolated and subjected to real-time PCR as described above.

Determination of the Set9 ternary complex with p53

The Set9 protein was prepared from a GST fusion as described previously¹⁷. A highly concentrated stock solution (100 mg ml⁻¹) of Set9 (aa108–366) was prepared in 50 mM Tris-HCl, pH 7.0, 100 mM NaCl, and diluted to 10 mg ml⁻¹ with a twofold molar excess of p53 10-mer peptide mono-methylated at Lys4 and AdoHcy. Crystals were grown at 18 °C by vapour diffusion as hanging drops prepared by mixing equal volumes of protein complex with a reservoir solution containing 0.1 M Tris-HCl, pH 7.8 and 22% PEG 3350. Crystals were first transferred into mother liquor augmented with an additional 5% PEG 400, before plunging into liquid nitrogen. Data were collected from flash-cooled crystals at 100 K on an ADSC Q4R CCD detector at SRS Daresbury. Diffraction data were integrated and scaled using DENZO and SCALEPACK²⁰. The structure was solved by molecular replacement using our previous model (1o9s.brk) with AMORE. Subsequent refinement was performed using REFMAC version 5.0 (ref. 21) and manual model building in O (ref. 22).

Determination of p53 peptide affinities

The dissociation constants were determined using fluorometric competition assays. The p53 and H3 peptides were unlabelled and used to displace a dansyl-labelled H3 10-mer (ARTKQTARKY) of known affinity. The resulting displacement curves were used to calculate the *K_d* of the unlabelled peptides: H3 20-mer (ARTKQTARKSTGGKAPRKQY) and p53 WT 20-mer (LKSCKGQSTSRHKLMFKTY). These peptides were tested against the Set9 construct consisting of residues 52–366. Fluorescence measurements were made using a Spex Fluoromax spectrophotometer with 330 nm (excitation) and 520 nm (emission) wavelengths respectively. The titrations were performed at 20 °C in a buffer containing 25 mM Tris-HCl pH 8.0.

Received 17 August; accepted 13 October 2004; doi:10.1038/nature03117.

Published online 3 November 2004.

- Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
- Wang, H. *et al.* Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase. *Mol. Cell* **8**, 1207–1217 (2001).
- Nishioka, K. *et al.* Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes Dev.* **16**, 479–489 (2002).
- Appella, E. & Anderson, C. W. Post-translational modifications and activation of p53 by genotoxic stresses. *Eur. J. Biochem.* **268**, 2764–2772 (2001).
- Shieh, S. Y., Ikeda, M., Taya, Y. & Prives, C. DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell* **91**, 325–334 (1997).
- Momand, J., Wu, H. H. & Dasgupta, G. MDM2—master regulator of the p53 tumor suppressor protein. *Gene* **242**, 15–29 (2000).
- Brooks, C. L. & Gu, W. Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr. Opin. Cell Biol.* **15**, 164–171 (2003).
- Li, M., Luo, J., Brooks, C. L. & Gu, W. Acetylation of p53 inhibits its ubiquitination by Mdm2. *J. Biol. Chem.* **277**, 50607–50611 (2002).
- Luo, J. *et al.* Acetylation of p53 augments its site-specific DNA binding both *in vitro* and *in vivo*. *Proc. Natl Acad. Sci. USA* **101**, 2259–2264 (2004).
- Gu, W. & Roeder, R. G. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell* **90**, 595–606 (1997).
- Barlev, N. A. *et al.* Acetylation of p53 activates transcription through recruitment of coactivators/histone acetyltransferases. *Mol. Cell* **8**, 1243–1254 (2001).
- Gu, W., Luo, J., Brooks, C. L., Nikolaev, A. Y. & Li, M. Dynamics of the p53 acetylation pathway. *Nov. Found. Symp.* **259**, 197–205 (2004).
- Gostissa, M. *et al.* Activation of p53 by conjugation to the ubiquitin-like protein SUMO-1. *Embo J.* **18**, 6462–6471 (1999).
- Kwek, S. S., Derry, J., Tyner, A. L., Shen, Z. & Gudkov, A. V. Functional analysis and intracellular localization of p53 modified by SUMO-1. *Oncogene* **20**, 2587–2599 (2001).
- Rodriguez, M. S. *et al.* SUMO-1 modification activates the transcriptional response of p53. *Embo J.* **18**, 6455–6461 (1999).

16. Kouskouti, A., Scheer, E., Staub, A., Tora, L. & Talianidis, I. Gene-specific modulation of TAF10 function by SET9-mediated methylation. *Mol. Cell* **14**, 175–182 (2004).
17. Xiao, B. *et al.* Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature* **421**, 652–656 (2003).
18. Grand, R. J. *et al.* The high levels of p53 present in adenovirus early region 1-transformed human cells do not cause up-regulation of MDM2 expression. *Virology* **210**, 323–334 (1995).
19. Tachibana, M., Sugimoto, K., Fukushima, T. & Shinkai, Y. SET domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *J. Biol. Chem.* **276**, 25309–25317 (2001).
20. Otwinowski, Z. & Minor, W. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 552–562 (SERC Daresbury Laboratory, Warrington, 1993).
21. CCP4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
22. Jones, T. A., Zhou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Vales for comments on the manuscript. We also thank S. L. Berger and members of the Reinberg laboratory, especially A. Kuzmichev and K. Sarma for discussions. We also thank E. White for discussions and P. Chumakov for the gift of the lentivirus LSL–GFP vector, and A. Ivanov for generation of the siRNA–Set9 U2OS cell line. We thank S. Martin for assistance with binding experiments. The GRASP centre at Tufts is acknowledged for providing excellent technical support. This work was supported by a grant from the NIH and the Howard Hughes Medical Institute (D.R.). N.A.B. acknowledges support from the Charlton Award. S.J.G. acknowledges support from the Medical Research Council and from the Association for International Cancer Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.R. (reinbedf@UMDNJ.EDU) or N.A.B. (nbarlev@tufts-nemc.org). Coordinates have been deposited in the Protein Data Bank under accession code 1XQH.

Luminal gating mechanism revealed in calcium pump crystal structures with phosphate analogues

Chikashi Toyoshima, Hiromi Nomura* & Takeo Tsuda

¹Institute of Molecular and Cellular Biosciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-0032, Japan

* Present address: Department of Information Physiology, National Institute for Physiological Sciences, Okazaki 444-8787, Japan

P-type ion transporting ATPases are ATP-powered ion pumps that establish ion concentration gradients across biological membranes. Transfer of bound cations to the luminal or extracellular side occurs while the ATPase is phosphorylated. Here we report at 2.3 Å resolution the structure of the calcium-ATPase of skeletal muscle sarcoplasmic reticulum, a representative P-type ATPase that is crystallized in the absence of Ca^{2+} but in the presence of magnesium fluoride, a stable phosphate analogue. This and other crystal structures determined previously provide atomic models for all four principal states in the reaction cycle. These structures show that the three cytoplasmic domains rearrange to move six out of ten transmembrane helices, thereby changing the affinity of the Ca^{2+} -binding sites and the gating of the ion pathway. Release of ADP triggers the opening of the luminal gate and release of phosphate its closure, effected mainly through movement of the A-domain, the actuator of transmembrane gates.

Cation pumps that establish ion gradients across biological membranes belong to the P-type ATPase superfamily. This includes Na^+K^+ -ATPase, gastric H^+K^+ -ATPase and other important ion pumps (see refs 1 and 2 for reviews). According to the classical E1/E2 theory^{3–5}, active transport by P-type ATPases is achieved by alternating the affinity and accessibility of the transmembrane ion binding sites: they have high affinity for the activating ion in E1 and low affinity in E2; they face the cytoplasm in E1 and the lumen or extracellular medium in E2. Actual transfer of the activating ion takes place while the ATPase is phosphorylated at an aspartyl residue. That is, bound ions occluded in the transmembrane binding sites in E1P-ADP and E1P are released into the lumen (or outside of the cell) during the transition to E2P. At the same time, the ADP-sensitivity, that is, the ability of the phosphoenzyme to synthesize ATP from ADP, is lost. Hydrolysis of the aspartyl-phosphate completes the cycle.

In the P-type ATPase family, Ca^{2+} -ATPase of skeletal muscle sarcoplasmic reticulum (SERCA1a) is structurally and functionally the best studied member. It consists of a single polypeptide of 994 amino acid residues^{6,7}, and comprises three cytoplasmic domains designated as A (actuator), N (nucleotide) and P (phosphorylation) domains and 10 transmembrane helices⁸. We have previously determined its crystal structures in three states (reviewed in ref. 9). They are a Ca^{2+} -bound form (E1·2 Ca^{2+} , Protein Data Bank (PDB) accession code 1SU4)⁸, a Ca^{2+} -unbound form stabilized by a very potent inhibitor thapsigargin (TG) (E2(TG), PDB accession code 1IWO)¹⁰, and a Ca^{2+} , Mg^{2+} and adenosine 5'-(β,γ -methylene)triphosphate (AMPPCP)-bound form (E1·AMPPCP, PDB accession code 1VFP)¹¹. In the E1·AMPPCP structure, also solved by Sørensen *et al.*¹² (PDB accession code 1T5S), the γ -phosphate of AMPPCP, a non-hydrolysable analogue of ATP, is bound to Asp 351, the residue to be phosphorylated. Sørensen *et al.*¹² also determined an E1·AlF₄·ADP structure (PDB accession code 1T5T), which probably mimics an E1P·ADP state with AlF₄ as a stable phosphate analogue¹³. The structures of E1·AMPPCP and E1·AlF₄·ADP are virtually the same, with the bound Ca^{2+} occluded in the transmembrane binding sites^{11,12}.

Here we describe a crystal structure of the Ca^{2+} -ATPase with a

bound phosphate analogue¹⁴, MgF_4^{2-} , in the absence of Ca^{2+} at 2.3 Å resolution. This structure, abbreviated as E2· MgF_4^{2-} , contains Mg^{2+} next to the phosphate analogue and is stabilized with TG, as in the E2 crystals¹⁰. The structure appears to represent a state E2·P_i (refs 15, 16), just after the hydrolysis of the aspartylphosphate but before the release of phosphate from the ATPase. We compare this structure with our own model of E1·AlF₄·ADP at 2.7 Å resolution and that of E2(TG)¹⁰ at 3.1 Å resolution (Fig. 1).

Thus, the structures described here are not exact analogues of E1P and E2P, but have very close domain organizations as revealed by limited proteolysis^{15,17}. They show how the affinity of the transmembrane Ca^{2+} binding sites is altered, and how the luminal gate is opened and closed by events that occur around the phosphorylation site some 50 Å away. A fairly complete description of the entire transport cycle is now possible. We show that the orientation of the A-domain plays the key role in regulating the transmembrane gates.

Structure determination

Crystals of E1·AlF₄·ADP complex were formed only with a C2 symmetry and showed very similar unit cell parameters to those of E1·AMPPCP crystals¹¹. The structure was determined by generalized molecular replacement¹⁸ from the model built for E1·AMPPCP and refined at 2.7 Å resolution.

Ca^{2+} -ATPase forms a very stable complex^{14,19} with Mg^{2+} and F^- in the absence of Ca^{2+} . Crystals of two different symmetries formed but only in the presence of thapsigargin. The crystals of C2 symmetry diffracted to 2.9 Å resolution. An initial rough atomic model was built without much difficulty using generalized molecular replacement¹⁸, starting from the three cytoplasmic domains of E1·2 Ca^{2+} fitted into a low-resolution (8 Å) map⁸ derived by electron microscopy for vanadate-induced tubular crystals²⁰. This model was used for molecular replacement with the crystals of P2₁ symmetry, which contained four molecules in the asymmetric unit. The atomic model was refined at 2.3 Å resolution to an R_{free} of 24.8%. At this resolution it was clear that the bound phosphate analogue is MgF_4^{2-} , rather than MgF_3^- as in a Rho protein²¹ (Supplementary Fig. 1). Because one Mg^{2+} is also bound between Asp 351 and Asp 703, the

stoichiometry of Mg^{2+} and F^- agrees well with biochemical measurements^{22,23}. The differences among the four molecules were small. During the course of preliminary model building, ADP, introduced in the purification procedure^{24,25}, was observed in the structure. ADP was therefore included in the crystallization buffer for P_2 crystals.

Overview of the structural changes

In all of the three structures compared here, the three cytoplasmic domains adopt overall compact forms. The arrangement in $\text{E}_2\cdot\text{MgF}_4^{2-}$, however, is distinctly different from that in $\text{E}_1\cdot\text{AlF}_4\cdot\text{ADP}$ (Fig. 1). The A-domain is $\sim 110^\circ$ rotated around an axis approximately perpendicular to the membrane (Fig. 2a, b) passing near

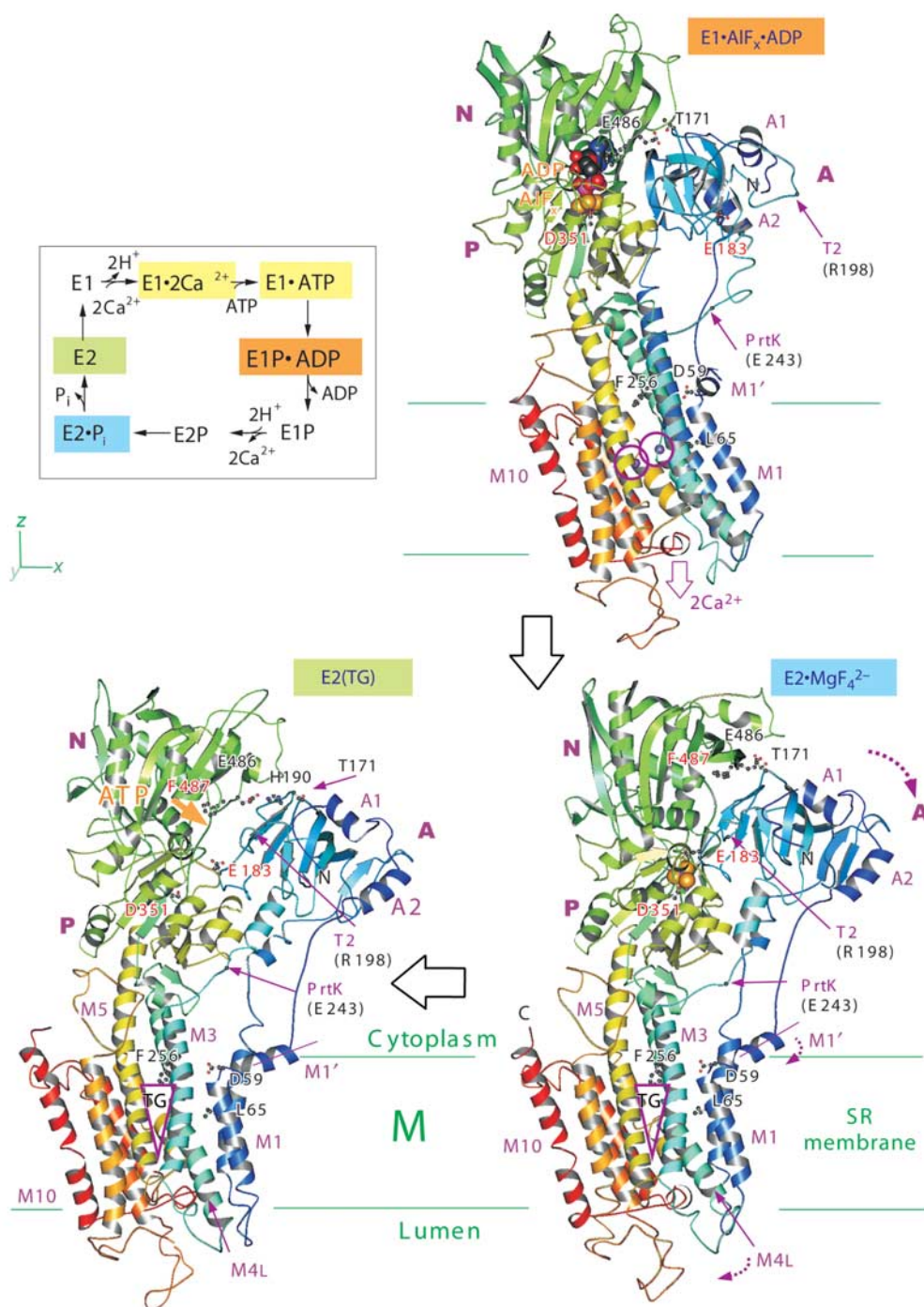


Figure 1 Front views (parallel to the membrane (x–y) plane) of Ca^{2+} -ATPase in three different states and a simplified reaction scheme (showing only the forward direction), in which different colours correspond to the respective structures presented here. The states whose names have yellow background refer to those previously described. In the ribbon models, colours change gradually from the amino terminus (blue) to the carboxy terminus (red). Purple spheres (circled) represent bound Ca^{2+} . Three cytoplasmic domains (A, N and P), the α -helices in the A-domain (A1 and A2) and those in the transmembrane domain (M1, M4L, M5 and M10) are indicated. M1' is an amphipathic part lying on the

bilayer surface. PrtK, a proteinase-K digestion site (around Glu 243 (ref. 45)); SR, sarcoplasmic reticulum; T2, a trypsin digestion site at Arg 198 (ref. 7); ATP, the binding pocket for the adenosine moiety of ATP; TG, thapsigargin. Several key residues—E183 (A), F487 (N, adenine binding), D351 (P, phosphorylation site), D59, L65 (M1) and F256 (M3, TG binding)—and those involved in interdomain hydrogen bonds (including T171, H190 and E486), are shown in ball-and-stick representation. Prepared with Molscript⁴⁶. Secondary structure was assigned with DSSP (ref. 47).

Val 705 in the P-domain, the residue forming a platform for tilting of the A-domain in the $E1\cdot 2Ca^{2+} \rightarrow E1\cdot AMPPCP$ transition. Tight association of the A- and P-domains is achieved through bound MgF_4^{2-} and Mg^{2+} (see later) and stabilized by several hydrogen bonds (including those between Arg 198 and Asp 660 and between Asp 203 and Arg 678 (Fig. 2c, inset)). Compared with $E1\cdot AlF_x\cdot ADP$, the N-domain associates with the P-domain very weakly, moved away from the P-domain by $\sim 60^\circ$. This is because the bridges between the N- and P-domains, facilitated by the ADP moiety of ATP^{11} , are lost and the A-domain wedges into the space. The interface between the N- and A-domains is also loose (Figs 1 and 2) and only two hydrogen bonds can be identified. The one between Thr 171 (A) and Glu 486 (N) is present also in $E1\cdot AlF_x\cdot ADP$, forming a mechanical couple (refer to Fig. 6 of ref. 11).

Reflecting these differences in the organization of cytoplasmic domains, the transmembrane helices M1–M6 are markedly rearranged between $E1\cdot AlF_x\cdot ADP$ and $E2\cdot MgF_4^{2-}$ (Figs 3a, b and 4a). The M3 and M4 helices are moved one turn of an α -helix towards the lumen (Fig. 3a, b); the M5 helix is bent towards M1 (Fig. 1), tilting the P-domain $\sim 30^\circ$ with respect to the membrane plane. These rearrangements, resulting in the release of Ca^{2+} into the lumen of the sarcoplasmic reticulum, are similar to those described previously for the $E1\cdot 2Ca^{2+} \rightarrow E2(TG)$ transition¹⁰.

This similarity arises because the arrangements of the cytoplasmic domains and transmembrane helices are similar in $E2\cdot MgF_4^{2-}$ and $E2(TG)^{10}$ (Fig. 1). For example, the path of the M5 helix and that of the cytoplasmic part of M4, which have principal roles in changing the affinity of the transmembrane Ca^{2+} -binding sites¹⁰, are virtually the same (Fig. 2d). Thus, the structure around the Ca^{2+} -binding sites hardly changes. The luminal half of the M4 helix (M4L), however, shows a $\sim 20^\circ$ difference in inclination (Figs 1, 3c and 4b), apparently closing the luminal gate in $E2(TG)$. This movement of M4L is related to that of the A-domain, which tilts $\sim 30^\circ$ with respect to the membrane plane (Figs 1 and 2d). This movement is caused by the release of Mg^{2+} and MgF_4^{2-} , which relaxes the P-domain and makes ineffective the secondary hinge in the β -sheet in the P-domain¹¹; as a result, the N-domain also changes its inclination ($\sim 30^\circ$), forming a different interface with the A-domain (Fig. 1).

Structure around the phosphorylation site

The most characteristic feature in $E2\cdot MgF_4^{2-}$ is a tight association of the A- and P-domains. The structure of the P-domain itself is virtually unaltered from $E1\cdot AMPPCP$ to $E2\cdot MgF_4^{2-}$. Given the high similarity in the P-domain structure between $E1\cdot 2Ca^{2+}$ and $E2(TG)^{10}$, this means that the P-domain of Ca^{2+} -ATPase assumes,

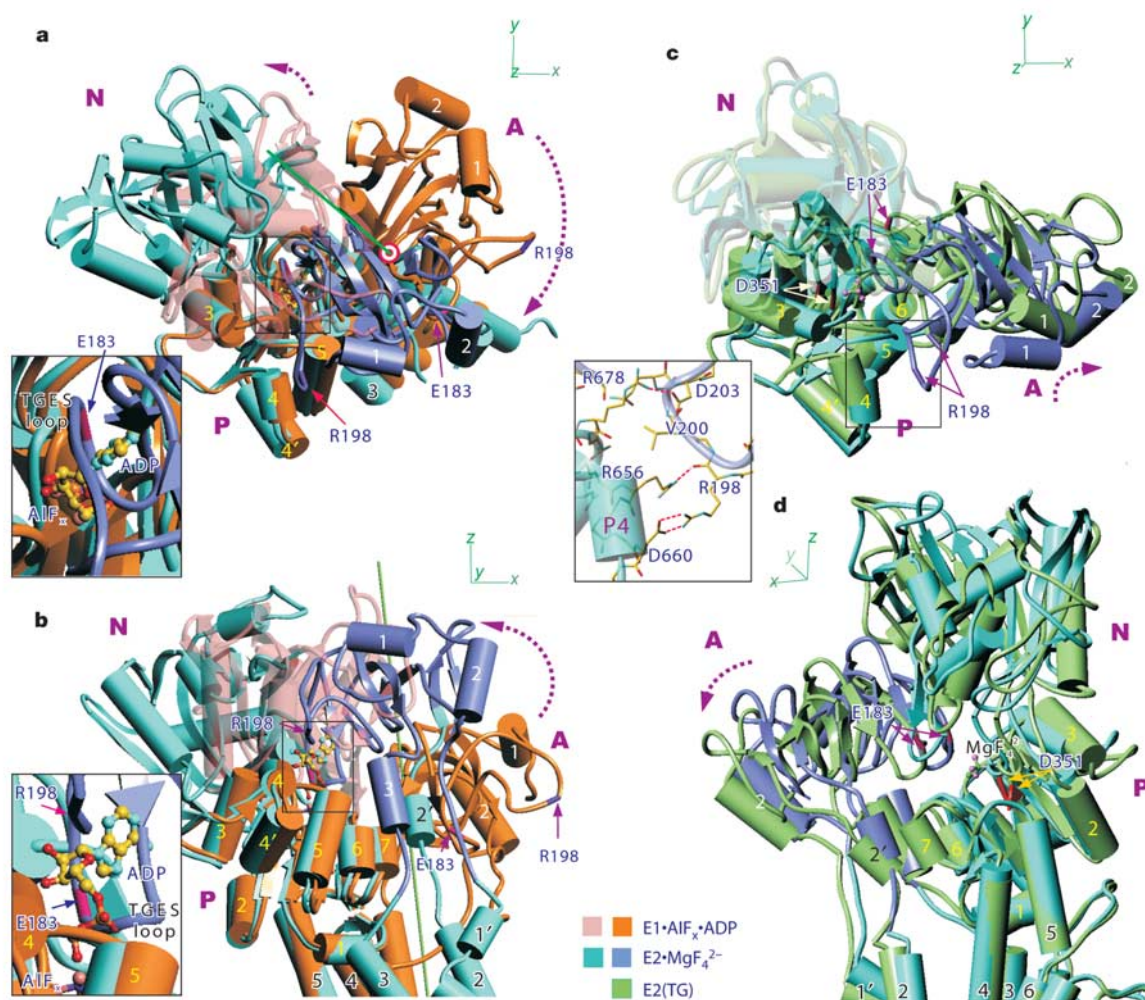


Figure 2 Movements of the cytoplasmic domains. The cytoplasmic domains of the $E2\cdot MgF_4^{2-}$ complex are superimposed with those of $E1\cdot AlF_x\cdot ADP$ (**a**, **b**) and $E2(TG)$ (**c**, **d**), aligned with the top part of the M5 helix integrated in the P-domain. Viewing directions are specified with x, y, z axes. Insets in **a** and **b** show the area around the phosphorylation site (Asp 351, coloured red in **c** and **d**); the one in **c** shows the interface between the A-

and P-domains around Arg 198. Ligands are shown in ball-and-stick representation. Different colours are used for the A-domain in $E2\cdot MgF_4^{2-}$ and the N-domain in $E1\cdot AlF_x\cdot ADP$ (transparent in **a** and **b**), to make the A-domain/N-domain interaction clearer. Large dotted arrows show the directions of expected movements in $E1P \rightarrow E2P$ and $E2P \rightarrow E2$.

as a first approximation, only two distinct structures depending on the phosphorylation state (or γ -phosphate binding). In $E2\cdot MgF_4^{2-}$, the TGES signature sequence starting from Thr 181 (ref. 26) in the A-domain now participates in the coordination of MgF_4^{2-} and

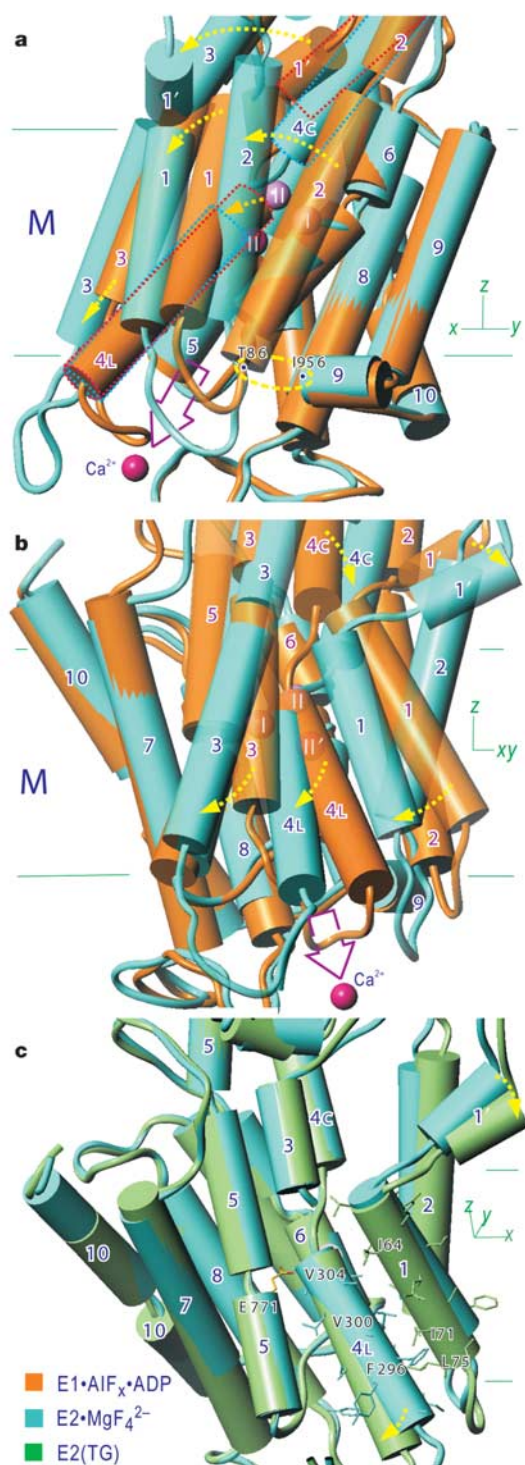


Figure 3 Movements in the transmembrane domain. The transmembrane domain of $E2\cdot MgF_4^{2-}$ is superimposed with $E1\cdot AlF_x\cdot ADP$ (a, b) and $E2(TG)$ (c), and viewed along the membrane plane. Transmembrane helices are numbered. Yellow arrows show the directions of movements in $E1\cdot AlF_x\cdot ADP \rightarrow E2\cdot MgF_4^{2-}$ (a, b) and $E2\cdot MgF_4^{2-} \rightarrow E2(TG)$ (c). Positions of Ca^{2+} in $E1\cdot AlF_x\cdot ADP$ are shown together with that expected (II') for site II Ca^{2+} after transition to $E2\cdot MgF_4^{2-}$ (a, b). Positions of the two halves of M4 (M4C and M4L) are outlined with dotted lines (a). Large open arrows indicate the release of Ca^{2+} into the lumen.

Mg^{2+} (Fig. 5), associating the A- and P-domains tightly. In $E1\cdot 2Ca^{2+}$ to $E1\cdot AlF_x\cdot ADP$, this sequence is located on the outermost surface of the ATPase, but brought into the phosphorylation site by the A-domain rotation (Figs 1 and 2a, b).

Of these four well-conserved and mutation-sensitive²⁷ residues in the TGES motif, Gly 182 and Glu 183 have clear functional roles. The carbonyl group of Gly 182 is located above the Mg^{2+} bound to Asp 703 and contributes to its binding through a water molecule (Fig. 5). This water molecule is also present in $E1\cdot AMPPCP^{11,12}$ and $E1\cdot AlF_x\cdot ADP^{12}$ but not fixed from above. The contribution of Gly 182 explains the higher affinity of the divalent metal at this site in $E2P$ than in $E1P^{28}$. Glu 183 has one of its carboxyl oxygens at a hydrogen bond distance from a fluorine in MgF_4^{2-} (Fig. 5). Considering the high electronegativity of fluorine, this configuration will be unstable unless that carboxyl oxygen is protonated. This in turn suggests that the carboxyl of Glu 183 is forced into this position by surrounding structure.

The importance of Glu 183 may be understood by considering the phosphorylated structure. As mentioned earlier, the structure of $E2\cdot MgF_4^{2-}$ probably represents that of $E2\cdot P_i$, immediately after the hydrolysis of the aspartylphosphate¹⁶. An atomic model of an aspartylphosphate can be taken from that of CheY (PDB accession code 1QMP)²⁹, a bacterial response regulator, which uses identical residues for the stabilization of phosphoryl group and Mg^{2+} despite the different protein folding pattern³⁰. In Fig. 5b, the model of aspartylphosphate is superimposed with the main chain to Asp 351. The oxygens of the phosphoryl group of the aspartylphosphate superimposed on the fluorines of MgF_4^{2-} without any change in the side-chain conformation of Asp 351. The aspartylphosphate does not have an oxygen atom corresponding to the distal fluorine, which is presumably hydrogen-bonded to Glu 183 in Ca^{2+} -ATPase. It is therefore likely that a water molecule normally occupies this position, at a distance of 2.8 Å from the carboxyl oxygen of Asp 351 (Fig. 5b), fixed by Glu 183 and the carbonyl of Thr 181, and makes an in-line attack on the aspartylphosphate to initiate hydrolysis (Fig. 5b). A mutagenesis study showed that Glu 183 is critical for $E2P$ hydrolysis and phosphorylation from P_i (ref. 31).

Then, accurate positioning of Glu 183 will be paramount. Its carboxyl group is fixed by hydrogen bonds with Thr 353, a crucial residue^{32,33}, and indirectly through a water molecule with Asp 601 in the hinge region (Fig. 5). Stabilization of the TGES loop itself through the Thr 181 hydroxyl group may also be important, as mutagenesis studies demonstrated²⁷. Hydrogen bonds between Ser 184 and Asn 359 in the hinge region, and between Ser 186 and Glu 439 in the N-domain, will also contribute. Furthermore, Gly 182 is bound to Mg^{2+} through a water molecule. This may be critical for decreasing the chance of a back reaction because the release of phosphate will release Mg^{2+} , which in turn will destabilize the TGES loop.

Roles of ADP-release

As described previously¹¹, there is a large space, presumably filled with bulk water, around the γ -phosphate of AMPPCP bound to Asp 351. This space is occupied by the A-domain in $E2\cdot MgF_4^{2-}$, in particular, with the hairpin containing the TGES signature sequence (Fig. 2a, b, inset). In fact, this hairpin occupies the space taken by ADP in $E1\cdot AlF_x\cdot ADP$. The TGES loop protrudes deeply into the P/N interface and probably even more so in genuine $E2P$ with the smaller aspartylphosphate (Fig. 5b).

The binding of ADP is allowed in $E2\cdot MgF_4^{2-}$ (Supplementary Fig. 2; ref. 23) because the binding pocket for the adenine ring (around Phe 487) in the N-domain is still available. Mg^{2+} appears to bridge α - and β -phosphates without ligating protein, as in $E1\cdot AlF_x\cdot ADP$. However, ADP is prevented from interacting with the phosphorylation site by the TGES loop (Supplementary Fig. 2) and synthesis of ATP in the reverse reaction is prohibited. This

feature explains the ADP-insensitivity of E2P. One of the differences of E2-MgF_4^{2-} from E2P is that the superfluorescence of TNP-AMP, characteristic of the nucleotide bound to E2P, is not observed with E2-MgF_4^{2-} (ref. 16). Deeper penetration of the TGES loop in genuine E2P could confine the space for nucleotide, and steric factors could play a part in the modulations of the reactions of E2P and E2-Pi by ATP^{34,35}.

Then the question arises: in the transition from E1P-ADP to E2P, does the TGES loop directly remove the ADP from the binding site? This seems unlikely, because the large rotation of the A-domain is only possible after the relaxation of the N-domain-P-domain association. In E1-AMPPCP and E1- $\text{AlF}_x\text{-ADP}$, β -phosphate seems to be critical for keeping the N-domain in the highly inclined position¹¹. Thus, the trigger for the transition into E2P is likely to be the release of ADP.

Release of Ca^{2+} into the lumen

In the E1P $\rightarrow$ E2P transition, the transmembrane helices undergo marked rearrangement. This rearrangement is essentially the same as in $\text{E1-2Ca}^{2+} \rightarrow \text{E2(TG)}^{10}$ in that the movements of the M3 and M4 helices have large downward components perpendicular to the membrane plane (Fig. 3a, b) apparently pushing out the Ca^{2+} , and in that the M5 helix bends substantially towards M1 (Fig. 1). These movements are mechanically linked, as described previously¹⁰, and decrease the number of oxygen atoms that can participate in the coordination of Ca^{2+} , thereby lowering the affinity for Ca^{2+} . The geometry of coordinating residues changes more markedly at site II (ref. 10), in which main-chain carbonyls of Val 304, Ala 305, Ile 307 (M4) and side-chain oxygens of Asn 796, Asp 800 (M6) and Glu 309 (M4) contribute in E1-2Ca^{2+} (ref. 8).

There is a critical difference between E2-MgF_4^{2-} and E2(TG) , however, relating to the release of Ca^{2+} , because the $\text{E1-2Ca}^{2+} \rightarrow \text{E2}$ transition (backward reaction) releases Ca^{2+} into the cytoplasm whereas the transition into E2P (forward reaction) releases Ca^{2+} into the lumen by opening the luminal gate. The structure around the transmembrane Ca^{2+} -binding sites is depicted in Fig. 4, in which the positions of the Ca^{2+} in E1- $\text{AlF}_x\text{-ADP}$ (I and II, violet spheres) are shown together with the expected position of site II

Ca^{2+} immediately after the transition into E2P (II', red sphere). In E1- $\text{AlF}_x\text{-ADP}$, the luminal half of the M4 helix (M4L, orange tube) comes closest to M6 near the Ca^{2+} at binding site II. Although displaced, site II Ca^{2+} is still almost on M4L, coordinated by the main-chain carbonyl of Val 304. In E2-MgF_4^{2-} (blue structure), M4L becomes parallel to M6, and forms a smooth interface devoid of bulky residues (Fig. 4b). If the site II Ca^{2+} moves together with Glu 309, which is suggested from the similar conformation in either state, it will come into the space between M4 and M6. This is because M4L changes its orientation relative to the loop containing Glu 309 (Figs 3b and 4a); the cytoplasmic part of M4 (M4C) is pushed towards M1 by M5 (ref. 10) whereas M4L is pushed by M1 in a different direction (Figs 3b and 4a).

Here the Ca^{2+} (II') meets Val 304, a well-conserved residue, Ala 301 or Leu 792 and must pass through the space surrounded by the M4 to M6 helices to reach the lumen (Fig. 4b; Supplementary Fig. 3). The side-chain nitrogen of Asn 796 comes close to the Ca^{2+} . This disposition, stabilized by hydrogen bonds with Glu 771, may help release Ca^{2+} into the lumen and block the entry of cations (other than H^+ to be countertransported) from the luminal side. It is interesting that this residue is aspartic acid in Na^+K^+ - and H^+K^+ -ATPases, which countertransport K^+ rather than H^+ .

We must remember, however, that the luminal gate is probably not fully open in E2-MgF_4^{2-} : a high concentration of Ca^{2+} (>mM) and a long time (>min) are required to reactivate the pump¹⁴. Also, the fluorescence from the tryptophans in the transmembrane region of E2-MgF_4^{2-} is largely diminished to a level similar to E2 (ref. 16). In genuine E2P, the passage is presumably wider. One way for predicting the structure in genuine E2P will be an extrapolation of the movement in $\text{E2(TG)} \rightarrow \text{E2-MgF}_4^{2-}$. In $\text{E2-MgF}_4^{2-} \rightarrow \text{E2(TG)}$, as described earlier, the release of the A-domain causes a movement of M4L (Figs 1, 3c and 4b) and decreases the space surrounded by M4-M6 helices (Fig. 4b). Thus, the exit channel is closed. In genuine E2P, the A-domain will be more tilted than in E2-MgF_4^{2-} towards the phosphorylation site and so will M1-M2, locating the luminal end of M4L further away from M5 and M6. In such a position, the space surrounded by M4-M6 will be larger and will allow Ca^{2+} to pass through.

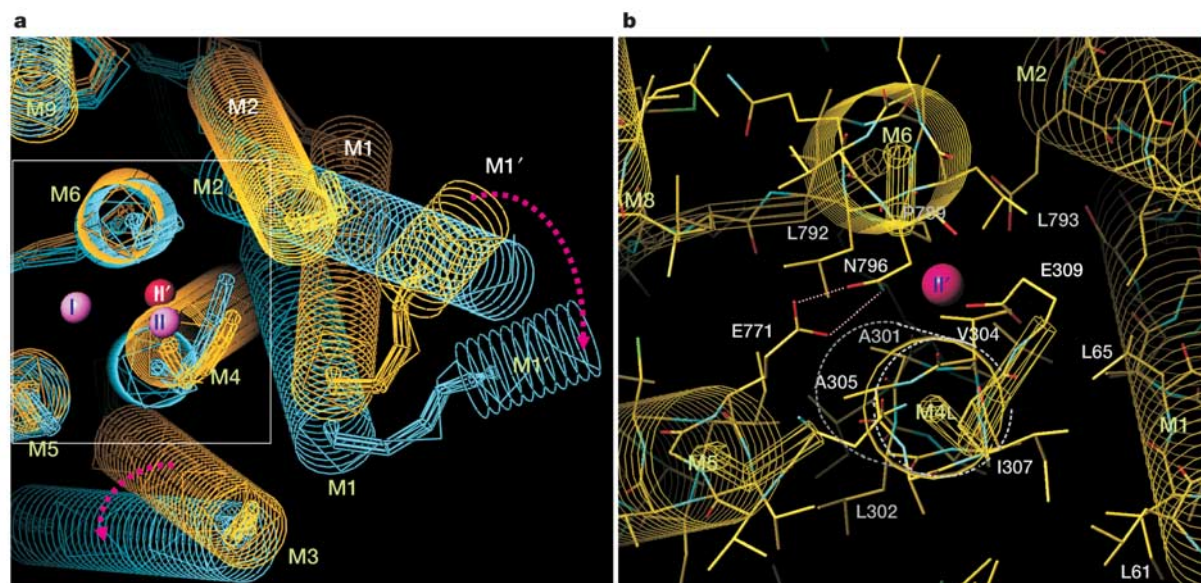


Figure 4 Movements in the transmembrane domain viewed approximately normal to the membrane plane from the cytoplasmic side. Superimposition of the transmembrane domains of E2-MgF_4^{2-} and E1- $\text{AlF}_x\text{-ADP}$ (**a**) and details in E2-MgF_4^{2-} (**b**) of the boxed area in **a**. Spheres in violet show the Ca^{2+} in E1- $\text{AlF}_x\text{-ADP}$ and those in red the expected

position of Ca^{2+} after transition to E2-MgF_4^{2-} . Arrows in **a** show the movements of transmembrane helices in $\text{E1-AlF}_x\text{-ADP} \rightarrow \text{E2-MgF}_4^{2-}$. The cylinder outlined by a dashed line in **b** outlines the M4L in E2(TG).

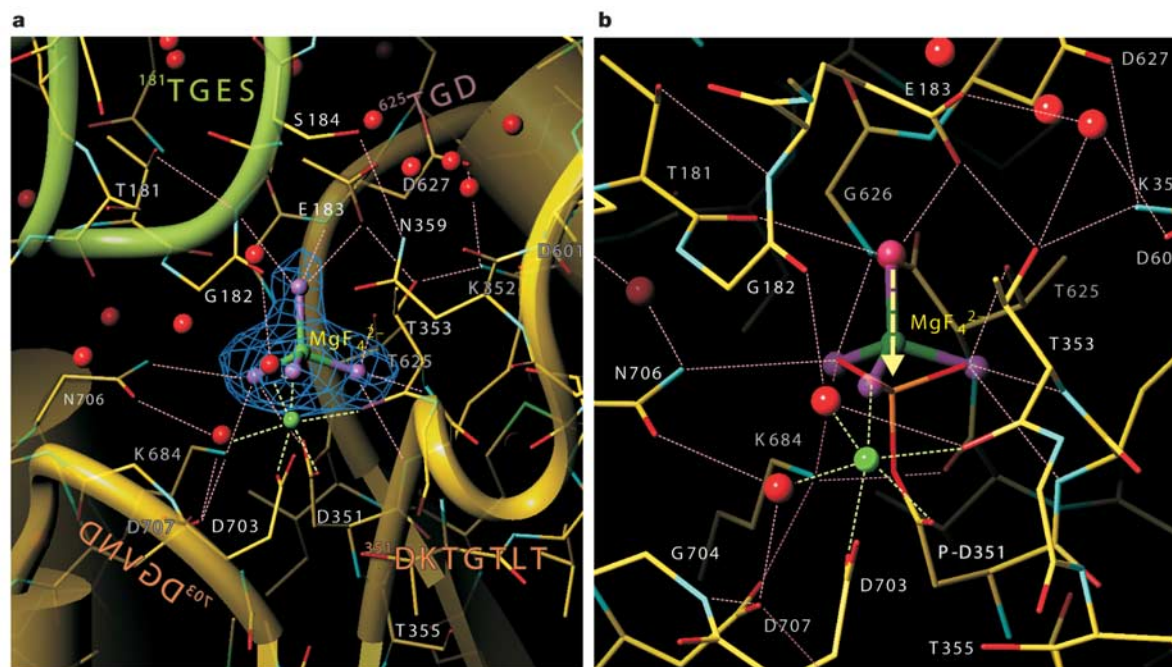


Figure 5 Details of the phosphorylation site in $E2\text{-MgF}_4^{2-}$. In the enlarged view (**b**), the atomic model of aspartylphosphate taken unchanged from a related protein CheY (PDB accession code 1QMP)²⁹ is incorporated. The blue net in **a** shows an omit annealed Fo–Fc map (at 5 σ ; temperature factor also refined) at 2.3 Å resolution. Small spheres represent

water molecules (red) and Mg^{2+} (green). MgF_4^{2-} is shown in ball-and-stick representation. Large yellow arrow indicates the expected water attack to the aspartylphosphate (P-D351). Conserved sequence motifs are shown in **a**. Broken lines in pink indicate likely hydrogen bonds, and those in light green coordinations of Mg^{2+} .

Links between the A-domain and transmembrane helices

Large movements of the transmembrane helices are linked to even larger rearrangements of the cytoplasmic domains. The rotation of the A-domain is likely to be triggered by the strain imposed on the M3–A-domain link, which is stretched by more than 3 Å (for Ala 240–Thr 247) in $E1\text{-AlF}_x\text{-ADP}$. The rotation appears to be stabilized by electrostatic interactions involving the Arg 198–Asp 202 loop (A-domain) and several charged residues in the P- and N-domains (Fig. 2c, inset). Positioning of this A-domain loop must be critical because mutations to Val 200 in the middle of this loop, which appears to orient long side chains of arginine and glutamic acid correctly and maintain them at optimal distances (Fig. 2c, inset), almost completely abolish ATPase activity³⁶.

The next question to ask is how the A-domain rotation is transmitted to the transmembrane helices. One interesting feature here is that the V-shaped structure formed by M1 (including M1') and M2 helices move as a rigid body. M1 has extensive van der Waals contacts with M2 and M4L (Fig. 3). Although M4L and M1 do not move as a rigid body, they keep the same van der Waals contacts from $E1\text{-AMPPCP}$ to $E2(\text{TG})$. Because M2 is a long continuous helix in $E1\text{-}2\text{Ca}^{2+}$ and pulled towards the A-domain in $E1\text{-}2\text{Ca}^{2+} \rightarrow E1\text{-AMPPCP}$, M2 is expected to have a firm interaction with the A-domain suitable for transmitting the movement of the A-domain to the transmembrane part. Then M1 will be able to move M4L with van der Waals contacts. In fact, the movement of the M1–M2 helices between $E1\text{-AlF}_x\text{-ADP}$ and $E2\text{-MgF}_4^{2-}$ can be regarded as a tilt of the V-shaped structure with the pivoting point around Thr 86 (at the luminal end of M2) and Ile 956 (on a short helix connecting M9 and M10 helices; Fig. 3a).

Rotation of the A-domain will tilt the M1–M2 helices, because M1 is already in contact with M4L, which cannot move much towards M5 or M6 due to steric collisions (Fig. 4a). Hence, the directions of the movement of M1 and M4 are similar (Figs 3a, b and 4a) but modified by their respective steric constraints. Because this is a tilting movement bringing the V-shaped structure more upright

(Fig. 3a), the movement has a component normal to the membrane (3.5 Å at Asp 59 on M1; 3.3 Å at Pro 308 on M4; Fig. 3a, b). In the $E2\text{-MgF}_4^{2-} \rightarrow E2(\text{TG})$ transition, it is easier to see the correlation between the movements of the V-shaped structure and M4L (Fig. 3c).

Limited proteolysis and mutagenesis studies also provide some

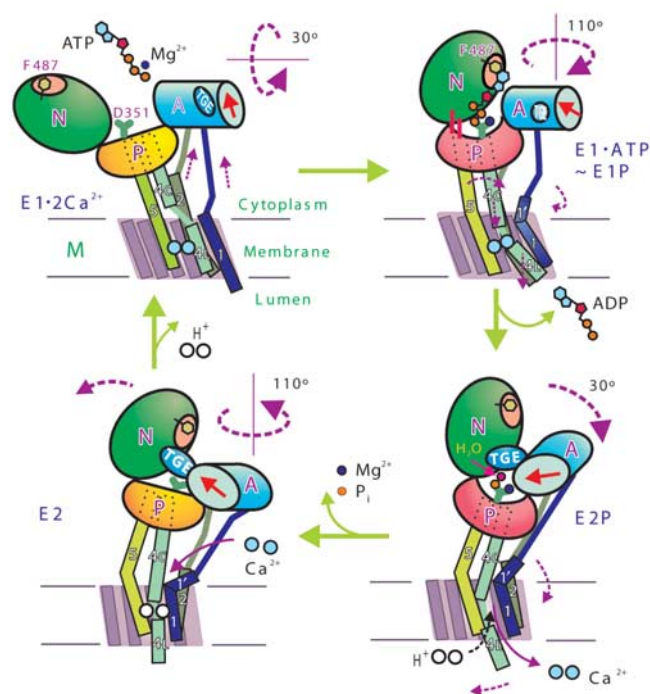


Figure 6 A cartoon depicting the structural changes of the Ca^{2+} -ATPase during the reaction cycle, based on the crystal structures in five different states.

clues. A nick at Lys 120 located at the C-terminal end of M2 substantially slows the $E2P \rightarrow E2$ transition but not $E1P \rightarrow E2P$ (ref. 37). Likewise, mutations of Tyr 122 block the $E2P \rightarrow E2$ transition but not $E1P \rightarrow E2P$ (ref. 38). Thus, the route through M2 appears critical in $E2P \rightarrow E2$ transition and less so in $E1P \rightarrow E2P$.

In $E1P \rightarrow E2P$, the connection to the A-domain through M3 seems to be more important, because a cleavage at Thr 242 on the loop (Fig. 1) strongly slows this transition³⁹. In addition, the top part of M3 seems to be strained in $E1\text{-AlF}_x\text{-ADP}$, bent towards M2¹¹, but not in $E2\text{-MgF}_4^{2-}$. This strain will be removed if the P-domain inclines and bends the M5 helix towards M1. These are certainly the movements required for releasing bound Ca^{2+} to the lumen. Mutagenesis studies suggest that links through M2 as well as the P-domain are also important. Arg 324 and Arg 334 on M4 are critical in the $E1P \rightarrow E2P$ transition³⁸. Arg 324 forms hydrogen bonds with Asn 101 and Gln 108 on M2 and Arg 334 with the P-domain in $E1\text{-AlF}_x\text{-ADP}$. Thus, for concerted movements of different parts, all these mechanical links seem to come into play.

Discussion

We now have five crystal structures that represent different states of the reaction cycle of Ca^{2+} -ATPase. They allow us to propose a synopsis of ion pumping (Fig. 6; Supplementary Movie).

- (1) Ca^{2+} binding to E2, which is the ground state, straightens the M5 helix and breaks the closed configuration of the three cytoplasmic domains, exposing the catalytic site. Two Ca^{2+} are bound in the high affinity sites formed by transmembrane helices M4, M5, M6 and M8 (ref. 8). The cytoplasmic gate is open and bound Ca^{2+} exchange with those in the cytoplasm⁴⁰. The M1 helix is deeply embedded in the lipid bilayer, stabilized indirectly by the bound Ca^{2+} .
- (2) ATP binds and crosslinks the P- and N-domains, so that the γ -phosphate of ATP and a Mg^{2+} bind to the P-domain to bend it¹¹. The N-domain is fixed in a highly inclined position and makes contact with the A-domain in a strained position. The M1 helix is pulled up and bent so that the top of the transmembrane part closes the cytoplasmic gate of the Ca^{2+} binding sites.
- (3) Phosphoryl transfer to Asp 351 allows the dissociation of ADP, which triggers the opening of the N- and P-domain interface; the A-domain rotates so that the TGES loop wedges into the gap and interacts with the phosphorylation site. This causes a marked rearrangement of the transmembrane helices M1–M6; large downward movements of M4, sharp bending of M5 and rotation of M6 destroy the Ca^{2+} -binding sites¹⁰. The lower sections of M1 and M2 push against M4L, opening the luminal gate and releasing the bound Ca^{2+} into the lumen.
- (4) The TGES loop of the A-domain fixes a particular water molecule and catalyses its attack on the aspartylphosphate. The release of the phosphate and Mg^{2+} unbends the P-domain. This in turn releases M1 and M2 so that M4L closes the luminal gate. The top amphipathic part of M1 (M1') forms a part of a cytoplasmic access funnel leading to Glu 309 (ref. 10), the gating residue of the Ca^{2+} binding sites⁴¹.

In short, the P- and N-domains change interfaces and thereby control the position of the A-domain, the actuator of transmembrane gates. ATP, phosphate, Mg^{2+} and Ca^{2+} are the modifiers of the interfaces. Energy barriers between the principal intermediate states seem to be comparable to the thermal energy, and we have seen devices integrated into Ca^{2+} -ATPase to decrease the chance of back reactions. Presumably the use of large-scale domain rearrangements, rather than small changes in the residues coordinating Ca^{2+} , is the means that nature has found to convert inherently stochastic thermal motions into concerted sequential movements with the help of ATP and the other modifiers. □

Methods

Crystallization

A MgF_4^{2-} complex of Ca^{2+} -ATPase was made by suspending sarcoplasmic reticulum vesicles in a buffer containing 1 mM KF, 10 mM MgCl_2 , 1 mM EGTA, 20% DMSO, 20% glycerol and 50 mM MOPS, pH 7.1. Ca^{2+} -ATPase was solubilized with 2% octaethyleneglycol mono-*n*-dodecylether (C_{12}E_8) and purified by affinity chromatography as described^{19,24,25}, but in the absence of Ca^{2+} . For elution from the column, 4 mM ADP was used. Affinity purified enzyme (20 μM) in C_{12}E_8 was mixed with thapsigargin (30 μM), then dialysed against a buffer consisting of 2.75 M glycerol, 10–12% PEG 400, 5 mM MgCl_2 , 3 mM KF, 0.06 mM ADP, 2.5 mM NaNa_3 , 2 $\mu\text{g ml}^{-1}$ butylhydroxytoluene, 0.2 mM dithiothreitol, 1 mM EGTA, 20 mM MES, pH 6.1, for about one month. Crystals were grown to $300 \times 300 \times 50 \mu\text{m}$ and flash-frozen in cold nitrogen gas.

Crystals of $\text{E1-AlF}_x\text{-ADP}$ were made by dialysing affinity purified enzyme against a buffer containing 2.75 M glycerol, 10–12% PEG 400, 2 mM AlCl_3 , 6 mM NaF, 0.3 mM ADP, 1 mM MgCl_2 , 0.2 M ammonium acetate, 3 mM CaCl_2 , 2.5 mM NaNa_3 , 0.2 mM dithiothreitol, 2 $\mu\text{g ml}^{-1}$ butylhydroxytoluene, 20 mM MES, pH 6.1. Crystals were grown to $200 \times 100 \times 50 \mu\text{m}$.

Data collection

Diffraction data were collected at BL41XU of SPring-8 using MAR-165 CCD detector and Rigaku R-Axis V imaging plate detector. A part of the initial data set for E2-MgF_4^{2-} was obtained at BL44XU using DIP2040 imaging plate detector. Diffraction intensities from the three best crystals were merged for E2-MgF_4^{2-} ($R_{\text{merge}} = 5.9\%$, $I/\sigma = 25.1$, redundancy = 5.1; 43.2%, 3.3, 2.0, respectively, for the highest resolution bin, 2.4–2.3 Å) and the two best crystals for $\text{E1-AlF}_x\text{-ADP}$ using Denzo and Scalepack⁴². Unit cell parameters were: $a = 102.2$, $b = 275.4$, $c = 109.9$ Å and $\beta = 90.01^\circ$ for E2-MgF_4^{2-} crystals of $P2_1$ symmetry; $a = 176.2$, $b = 70.0$, $c = 142.2$ Å and $\beta = 107.0^\circ$ for E2-MgF_4^{2-} crystals of $C2$ symmetry; $a = 163.1$, $b = 75.0$, $c = 151.5$ Å and $\beta = 109.3^\circ$ for $\text{E1-AlF}_x\text{-ADP}$ crystals.

Modelling

An initial model of E2-MgF_4^{2-} was built with the $C2$ crystals starting from the three cytoplasmic domains fitted into an 8 Å resolution map⁸ derived by electron microscopy of vanadate induced tubular crystals²⁰. PC-refinement¹⁸ converged immediately and the calculated map showed that the P-domain was significantly different from that in E1Ca^{2+} . Then, the rest of the model was built gradually by many cycles of map calculation, solvent flattening⁴³ and manual model building. (See Supplementary Methods for more details.)

The final model was made with the $P2_1$ crystals, which contain four molecules in the asymmetric unit. PC-refinement after dividing the model of a monomer into 10 segments showed that the two molecules in the asymmetric unit form a pair. The largest differences between the pairs were observed with the uppermost loop (around Arg 505, T1 trypsin site) and the luminal loop connecting M7 and M8 helices. The model includes four protomers, each of which contains Ca^{2+} -ATPase, TG, MgF_4^{2-} , 2Mg^{2+} and ADP, and altogether 594 water molecules. It was refined against the diffraction data consisting of 249,278 reflections (93.3% completeness) to R_{free} of 0.248 and R_{cryst} of 0.228 at 2.3 Å resolution; root mean square deviation (r. m. s. d.) of the bond length and angle were 0.007 Å and 1.2° , respectively.

The model of $\text{E1-AlF}_x\text{-ADP}$ was built similarly to that of E1-AMPPCP (ref. 11). Because the two structures were virtually the same, the model built for E1-AMPPCP was used as the starting model in the refinement. The model included Ca^{2+} -ATPase, ADP, AlF_x , 2Mg^{2+} and six water molecules. It was refined against the diffraction data consisting of 38,433 reflections. The modelling showed that both AlF_3 and AlF_4^- were possible with the map, as is the case with phosphoserine phosphatase⁴⁴, although the lack of electron density connecting the Mg^{2+} bound to ADP and the fourth fluorine in AlF_4^- (ref. 12) favours AlF_3 . The model including AlF_3 gave slightly better R_{free} (0.290; $R_{\text{cryst}} = 0.259$) than AlF_4^- ($R_{\text{free}} = 0.297$; $R_{\text{cryst}} = 0.258$) at 2.7 Å resolution; r. m. s. d. of the bond length and angle were 0.009 Å and 1.4° , respectively.

Received 31 July; accepted 1 September 2004; doi:10.1038/nature02981.

Published online 26 September 2004.

1. Möller, J. V., Juul, B. & le Maire, M. Structural organization, ion transport, and energy transduction of P-type ATPases. *Biochim. Biophys. Acta* **1286**, 1–51 (1996).
2. Kühlbrandt, W. Biology, structure and mechanism of P-type ATPases. *Nature Rev. Mol. Cell Biol.* **5**, 282–295 (2004).
3. Albers, R. W. Biochemical aspects of active transport. *Annu. Rev. Biochem.* **36**, 727–756 (1967).
4. Post, R. L., Hegyváry, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247**, 6530–6540 (1972).
5. de Meis, L. & Vianna, A. L. Energy interconversion by the Ca^{2+} -dependent ATPase of the sarcoplasmic reticulum. *Annu. Rev. Biochem.* **48**, 275–292 (1979).
6. Brandl, C. J., deLeon, S., Martin, D. R. & MacLennan, D. H. Adult forms of the Ca^{2+} ATPase of sarcoplasmic reticulum. Expression in developing skeletal muscle. *J. Biol. Chem.* **262**, 3768–3774 (1987).
7. MacLennan, D. H., Brandl, C. J., Korczak, B. & Green, N. M. Amino-acid sequence of a Ca^{2+} + Mg^{2+} -dependent ATPase from rabbit muscle sarcoplasmic reticulum, deduced from its complementary DNA sequence. *Nature* **316**, 696–700 (1985).
8. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**, 647–655 (2000).
9. Toyoshima, C. & Inesi, G. Structural basis of ion pumping by Ca^{2+} -ATPase of the sarcoplasmic reticulum. *Annu. Rev. Biochem.* **73**, 269–292 (2004).
10. Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418**, 605–611 (2002).

11. Toyoshima, C. & Mizutani, T. Crystal structure of the calcium pump with a bound ATP analogue. *Nature* **430**, 529–535 (2004).
12. Sørensen, T. L., Møller, J. V. & Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science* **304**, 1672–1675 (2004).
13. Troullier, A., Girardet, J. L. & Dupont, Y. Fluoroaluminate complexes are bifunctional analogues of phosphate in sarcoplasmic reticulum Ca^{2+} -ATPase. *J. Biol. Chem.* **267**, 22821–22829 (1992).
14. Murphy, A. J. & Coll, R. J. Fluoride is a slow, tight-binding inhibitor of the calcium ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **267**, 5229–5235 (1992).
15. Danko, S. *et al.* ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+} -ATPase has a compact conformation resistant to proteinase K, V8 protease and trypsin. *FEBS Lett.* **489**, 277–282 (2001).
16. Danko, S., Yamasaki, K., Daiho, T. & Suzuki, H. Distinct natures of beryllium fluoride-bound, aluminum fluoride-bound, and magnesium fluoride-bound stable analogues of an ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+} -ATPase: changes in catalytic and transport sites during phosphoenzyme hydrolysis. *J. Biol. Chem.* **279**, 14991–14998 (2004).
17. Danko, S., Yamasaki, K., Daiho, T., Suzuki, H. & Toyoshima, C. Organization of cytoplasmic domains of sarcoplasmic reticulum Ca^{2+} -ATPase in E₁P and E₁ATP states: a limited proteolysis study. *FEBS Lett.* **505**, 129–135 (2001).
18. Brünger, A. T. Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. *Acta Crystallogr. A* **46**, 46–57 (1990).
19. Yamasaki, K., Daiho, T. & Suzuki, H. Remarkable stability of solubilized and delipidated sarcoplasmic reticulum Ca^{2+} -ATPase with tightly bound fluoride and magnesium against detergent-induced denaturation. *J. Biol. Chem.* **277**, 13615–13619 (2002).
20. Zhang, P., Toyoshima, C., Yonekura, K., Green, N. M. & Stokes, D. L. Structure of the calcium pump from sarcoplasmic reticulum at 8-Å resolution. *Nature* **392**, 835–839 (1998).
21. Graham, D. L. *et al.* MgF_3^- as a transition state analog of phosphoryl transfer. *Chem. Biol.* **9**, 375–381 (2002).
22. Coll, R. J. & Murphy, A. J. Fluoride-inhibited calcium ATPase of sarcoplasmic reticulum. Magnesium and fluoride stoichiometry. *J. Biol. Chem.* **267**, 21584–21587 (1992).
23. Daiho, T., Kubota, T. & Kanazawa, T. Stoichiometry of tight binding of magnesium and fluoride to phosphorylation and high-affinity binding of ATP, vanadate, and calcium in the sarcoplasmic reticulum Ca^{2+} -ATPase. *Biochemistry* **32**, 10021–10026 (1993).
24. Coll, R. J. & Murphy, A. J. Purification of the CaATPase of sarcoplasmic reticulum by affinity chromatography. *J. Biol. Chem.* **259**, 14249–14254 (1984).
25. Stokes, D. L. & Green, N. M. Three-dimensional crystals of CaATPase from sarcoplasmic reticulum. Symmetry and molecular packing. *Biophys. J.* **57**, 1–14 (1990).
26. Palmgren, M. G. & Axelsen, K. B. Evolution of P-type ATPases. *Biochim. Biophys. Acta* **1365**, 37–45 (1998).
27. Clarke, D. M., Loo, T. W. & MacLennan, D. H. Functional consequences of mutations of conserved amino acids in the β -strand domain of the Ca^{2+} -ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **265**, 14088–14092 (1990).
28. Ogurusu, T., Wakabayashi, S. & Shigekawa, M. Activation of sarcoplasmic reticulum Ca^{2+} -ATPase by Mn^{2+} : a Mn^{2+} binding study. *J. Biochem.* **109**, 472–476 (1991).
29. Lewis, R. J., Brannigan, J. A., Muchova, K., Barak, I. & Wilkinson, A. J. Phosphorylated aspartate in the structure of a response regulator protein. *J. Mol. Biol.* **294**, 9–15 (1999).
30. Ridder, I. S. & Dijkstra, B. W. Identification of the Mg^{2+} -binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY. *Biochem. J.* **339**, 223–226 (1999).
31. Clausen, J. D., Vilsen, B., McIntosh, D. B., Einholm, A. P. & Andersen, J. P. Glutamate-183 in the conserved TGES motif of domain A of sarcoplasmic reticulum Ca^{2+} -ATPase assists in catalysis of E₂/E₂P partial reactions. *Proc. Natl Acad. Sci. USA* **101**, 2776–2781 (2004).
32. Maruyama, K. *et al.* Functional consequences of alterations to amino acids located in the catalytic center (isoleucine 348 to threonine 357) and nucleotide-binding domain of the Ca^{2+} -ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **264**, 13038–13042 (1989).
33. Clausen, J. D., McIntosh, D. B., Woolley, D. G. & Andersen, J. P. Importance of Thr-353 of the conserved phosphorylation loop of the sarcoplasmic reticulum Ca^{2+} -ATPase in MgATP binding and catalytic activity. *J. Biol. Chem.* **276**, 35741–35750 (2001).
34. McIntosh, D. B. & Boyer, P. D. Adenosine 5'-triphosphate modulation of catalytic intermediates of calcium ion activated adenosinetriphosphatase of sarcoplasmic reticulum subsequent to enzyme phosphorylation. *Biochemistry* **22**, 2867–2875 (1983).
35. Champell, P. *et al.* ATP regulation of sarcoplasmic reticulum Ca^{2+} -ATPase. Metal-free ATP and 8-bromo-ATP bind with high affinity to the catalytic site of phosphorylated ATPase and accelerate dephosphorylation. *J. Biol. Chem.* **263**, 12288–12294 (1988).
36. Kato, S. *et al.* Val 200 residue in Lys 189-Lys 205 outermost loop on the A domain of sarcoplasmic reticulum Ca^{2+} -ATPase is critical for rapid processing of phosphoenzyme intermediate after loss of ADP sensitivity. *J. Biol. Chem.* **278**, 9624–9629 (2003).
37. Lenoir, G. *et al.* Functional properties of sarcoplasmic reticulum Ca^{2+} -ATPase after proteolytic cleavage at Leu 119-Lys 120, close to the A-domain. *J. Biol. Chem.* **279**, 9156–9166 (2004).
38. Yamasaki, K., Daiho, T., Danko, S. & Suzuki, H. Multiple and distinct effects of mutations of Tyr 122, Glu 123, Arg 324, and Arg 334 involved in interactions between the top part of second and fourth transmembrane helices in sarcoplasmic reticulum Ca^{2+} -ATPase: changes in cytoplasmic domain organization during isometric transition of phosphoenzyme intermediate and subsequent Ca^{2+} release. *J. Biol. Chem.* **279**, 2202–2210 (2004).
39. Møller, J. V. *et al.* Calcium transport by sarcoplasmic reticulum Ca^{2+} -ATPase. Role of the A domain and its C-terminal link with the transmembrane region. *J. Biol. Chem.* **277**, 38647–38659 (2002).
40. Orlowski, S. & Champell, P. Kinetics of calcium dissociation from its high-affinity transport sites on sarcoplasmic reticulum ATPase. *Biochemistry* **30**, 352–361 (1991).
41. Inesi, G., Ma, H., Lewis, D. & Xu, C. Ca^{2+} occlusion and gating function of Glu 309 in the ADP-fluoroaluminate analog of the Ca^{2+} -ATPase phosphoenzyme intermediate. *J. Biol. Chem.* **279**, 31629–31637 (2004).
42. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–325 (1997).
43. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
44. Wang, W. *et al.* Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic “snapshots” of intermediate states. *J. Mol. Biol.* **319**, 421–431 (2002).
45. Juul, B. *et al.* Do transmembrane segments in proteolysed sarcoplasmic reticulum Ca^{2+} -ATPase retain their functional Ca^{2+} binding properties after removal of cytoplasmic fragments by proteinase K? *J. Biol. Chem.* **270**, 20123–20134 (1995).
46. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
47. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Suzuki for his contribution at the initial phase of this work. We acknowledge that the first crystals of $\text{E}_2\text{-MgF}_4^{2-}$ were made by Y. Tsubaki. Thanks are also due to M. Kawamoto, H. Sakai and E. Yamashita for data collection at SPring-8; N. Miyashita for making many movies; M. Takahashi and J. Tsueda for preparing figures; and Y. Ohuchi for computer programs. We are grateful to D. B. McIntosh for help in improving the manuscript and G. Inesi for communicating unpublished results to us. This work was supported in part by a Creative Science Project Grant from the Ministry of Education, Culture, Sports, Science and Technology, the Japan New Energy and Industry Technology Development Organization, and the Human Frontier Science Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.T. (ct@iam.u-tokyo.ac.jp). The atomic coordinates for E₁-AlF₆-ADP and E₂-MgF₄²⁻ are deposited in the PDB under accession codes 1WPE and 1WPG, respectively.

Molecular hydrogen beyond the optical edge of an isolated spiral galaxy

Jonathan Braine & Fabrice Herpin

Observatoire de Bordeaux, UMR 5804, CNRS/INSU, B.P. 89, F-33270 Floirac, France

Knowledge about the outermost portions of galaxies is limited owing to the small amount of light coming from them. It is known that in many cases atomic hydrogen (H I) extends well beyond the optical radius¹. In the centres of galaxies, however, molecular hydrogen (H₂) usually dominates by a large factor^{2–4}, raising the question of whether H₂ is also abundant in the outer regions. Here we report the detection of emission from carbon monoxide (CO), the most abundant tracer of H₂, beyond the optical radius of the nearby galaxy NGC 4414. The host molecular clouds probably formed in the regions of relatively high H I column density and in the absence of spiral density waves. The relative strength of the lines from the two lowest rotational levels indicates that both the temperature and density of the H₂ are quite low compared to conditions closer to the centre. The inferred surface density of the molecular material continues the monotonic decrease from the inner regions. We conclude that although molecular clouds can form in the outer region of this galaxy, there is little mass associated with them.

The CO spectra shown in Fig. 1 reveal the presence of cool, but not very cold, molecular gas out to 1.5 times the optical radius (R_{25}) of the isolated spiral NGC 4414. The CO observations were carried out in good conditions with the 30-m antenna on Pico Veleta (Spain) operated by the Institut de Radioastronomie Millimétrique, in May and November 2003, and March 2004. Cool H₂ is not directly observable because it has no permitted rotational transitions so CO, the most abundant heteronuclear molecular, is used as a proxy. The galaxy chosen is expected to be representative of a large class of spirals because it is of late type, quite axisymmetric⁵, and has no immediate neighbours with which it could interact. The atomic gas is extended, also indicating a lack of major interactions. This lack of interactions is important: although every interaction is different owing to the large number of parameters involved, isolated galaxies can reasonably be hoped to be representative. Furthermore, NGC 4414 is near the North Galactic Pole and suffers little from foreground (Galactic) confusion or extinction.

Almost no constraints are as yet available on the H₂ content of the outer disks of spirals. To our knowledge, these are the first published detections of CO beyond the optical radius of an external spiral, reaching $1.5R_{25}$. A. Ferguson also reports detection (personal communication) of CO in NGC 6946 beyond R_{25} . Digel *et al.*⁶

detected isolated clouds at large radii in the Milky Way, although the mass implied was quite low. Our previously published observations of NGC 4414, covering almost the whole optical disk, show a clear decline in the CO(2–1)/CO(1–0) line ratio (independent of beam size) which we interpret⁷ as a decrease in excitation temperature, meaning that even if substantial CO is present, the lines could be quite weak. Dust, which is necessary for H₂ formation, is clearly seen in absorption at the edge of the spiral disks of galaxies and even well beyond through the reddening of stars and background galaxies⁸. Observations of optical line ratios in H II regions at large radii also show that metals and star formation are present⁹.

Figure 1 summarizes the results for the positions (–8, 133), (–68, 162), (–98, 162), (31, –115), and (51, –138) in panels a to f. The main panel shows the galaxy NGC 4414 itself in R band with contours giving the H I column density. The circles indicate the beam size and the positions of the spectra shown individually. They are beyond the nominal optical radius R_{25} and extend up to $1.5R_{25}$ (–68, 162), with the position (–98, 162) not detected in CO.

The positions were chosen to maximize the likelihood of detecting CO by preferentially observing high H I column density regions. Indeed, the undetected position is quite close to a detected one but has only half the H I column density. This corroborates the findings of ref. 10, whose authors observed M33 with the BIMA interferometer in CO(1–0) and found molecular gas nearly exclusively in zones of high H I column density. They interpret this as showing that the H₂ forms from the high-density H I. M33 is a much smaller and lower metallicity spiral than NGC 4414, but there is also evidence for H I to H₂ conversion in the Milky Way^{11,12}. A further indication that the H₂ is forming from H I in the outer parts of NGC 4414 is that although the CO line profiles are always narrower than the H I, the linewidths seem to be correlated (compare in Fig. 1). If the H₂ is indeed forming from H I in NGC 4414, then the H₂ formation process does not require a passage through spiral arms, because no spiral structure is observed in NGC 4414 (ref. 5).

Using a ‘standard’ $N(\text{H}_2)/I_{\text{CO}(1-0)} = 2 \times 10^{20} \text{ H}_2 \text{ cm}^{-2} (\text{K km s}^{-1})^{-1}$ conversion ratio¹³, the H₂ column density is 5–10% of the H I, roughly $N_{\text{H I}} \approx 5 \times 10^{20} \text{ cm}^{-2}$. The linewidths of the spectra are greater than those of giant molecular clouds and the molecular gas mass implied is $(1-2) \times 10^6 M_\odot$, depending on the position. In the outer parts the $N(\text{H}_2)/I_{\text{CO}(1-0)}$ ratio is probably somewhat higher^{7,14} than the above-‘standard’ value which is used to enable comparison with previous work. It is commonly believed^{15–17} that the H I is a two-phase medium, with the warm diffuse phase becoming more and more dominant at increasing galactocentric distances. Clearly, our results show that even where the stellar mass (light) contribution is very low (dim), H₂ formation is still possible. Whether this extends further out to lower, and possibly warmer, H I columns remains to be seen and is certainly necessary to determine the H₂ mass of spirals.

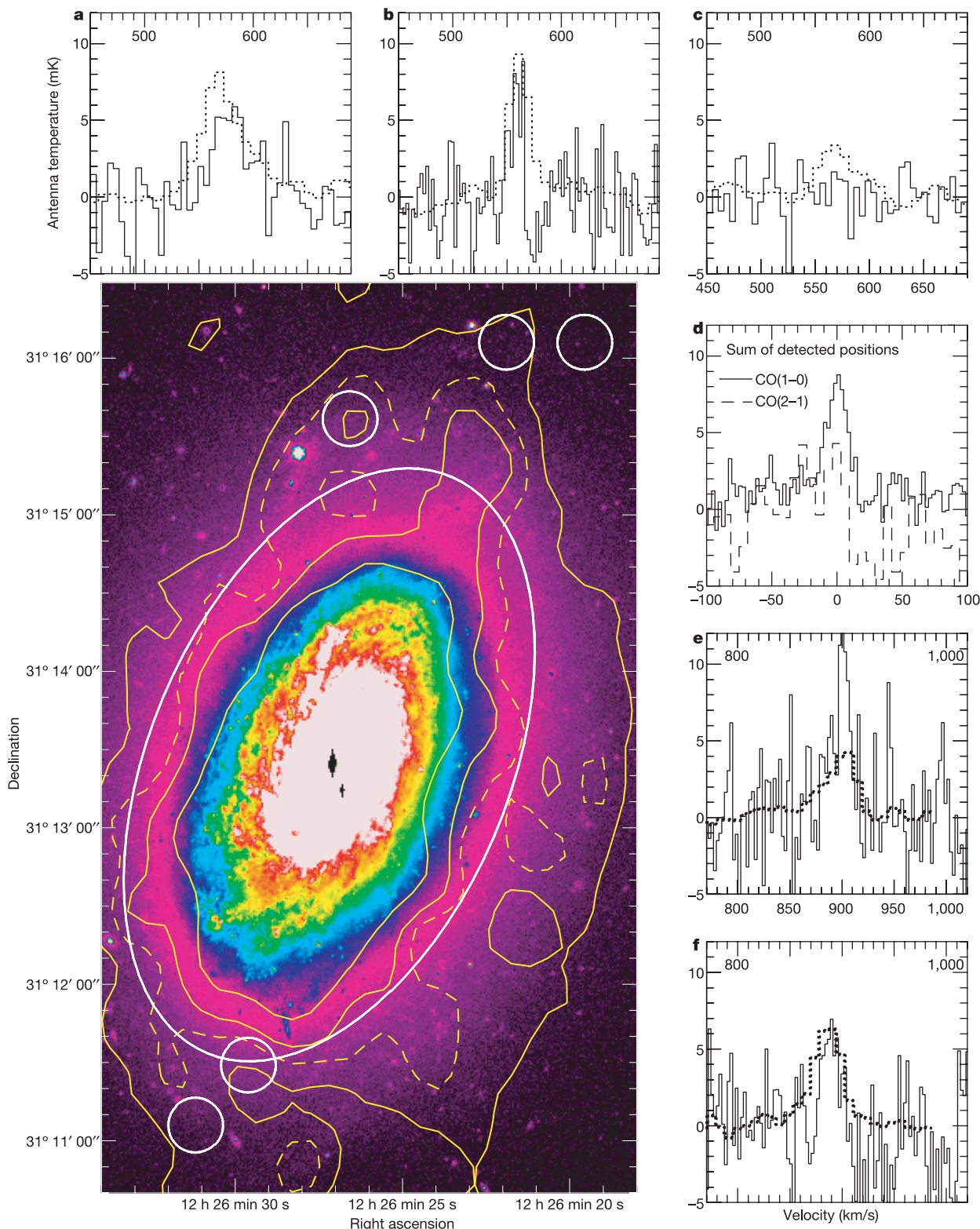
The CO detections presented here are really a continuous extension of the decline in CO brightness described in ref. 7, further

Figure 1 (on page 370) Spectra showing the carbon monoxide and atomic hydrogen emission from the outer regions of NGC 4414. An R-band image of NGC 4414 taken with the CFHT is shown with H I contours from recent Westerbork observations at column densities of 4, 6, 8 and $12 \times 10^{20} \text{ atoms cm}^{-2}$, the $6 \times 10^{20} \text{ atoms cm}^{-2}$ contour being dashed. The ellipse indicates the average R_{25} contour. The surrounding boxes show the CO (full line) and H I (dotted line) spectra for the positions indicated by the five circles. **a–c**, The positions (–8, 133), (–68, 162), and (–98, 162) are the top three circles. **e,f**, The positions (31, –115) and (51, –138) are the two lower circles. The CO(2–1) line was not clearly detected but the conditions were very good, enabling us to obtain a sensitive spectrum representing the sum of the four positions detected in CO(1–0). Panel **d** shows the CO(1–0) and CO(2–1) spectra summed over these positions. The line ratio is

$\text{CO}(2-1)/\text{CO}(1-0) \approx 0.5$ (0.4 from the integrated intensities and 0.5 from fitting gaussians). The CO(1–0) line spectrum has been offset by +1 mK and the CO(2–1) by –1 mK for clarity. The beam widths differ by a factor of two for the two transitions but only if the CO is concentrated systematically either towards the centre of the beam (better picked up in CO(2–1)) or about 5–10" away from the beam centre (that is, poorly picked up by the CO(2–1) beam) could the ratio be biased. Given that the linewidths are considerably greater than those of giant molecular clouds, the gas is not likely to be concentrated at any particular position with respect to the beam centre, rendering the ratio valid without further correction. All spectra are shown in the main beam temperature scale in millikelvins, the velocity being in kilometres per second (LSR, optical convention).

confirmed by unpublished CO observations at radii slightly less than R_{25} . Analysis of the H α image of ref. 18 shows that the H α brightness also decreases continuously through the R_{25} region, although the H I column density decreases rapidly¹⁹. It is generally believed—and confirmed for the Milky Way¹⁴ and NGC 4414⁷ through isotopic and dust continuum measurements—that the amount of H₂ is reasonably well known out to at least the solar

circle or equivalent. Thus, for H₂ to represent as much or more mass than the H I at R_{25} and beyond, the H₂ surface density must increase with radius over this range despite the regular decline in CO and H α brightness and the (known) decrease in H₂ and H I surface densities at lower radii. From an analysis of the CO line ratio in these outer positions, we estimate the H₂ to represent at most 30% of the H I surface density at $1.5R_{25}$; if H₂ were the dark matter in spiral



galaxies²⁰, 50 times as much H₂ would be required to fit the rotation curve. □

Received 18 June; accepted 21 September 2004; doi:10.1038/nature03054.

1. Casertano, S. & van Gorkom, J. H. Declining rotation curves—the end of a conspiracy? *Astron. J.* **101**, 1231–1241 (1991).
2. Young, J. S. & Scoville, N. Z. Molecular gas in galaxies. *Annu. Rev. Astron. Astrophys.* **29**, 581–625 (1991).
3. Combes, F. Distribution of CO in the Milky Way. *Annu. Rev. Astron. Astrophys.* **29**, 195–237 (1991).
4. Braine, J. *et al.* A CO(1–0) and CO(2–1) survey of nearby spiral galaxies. I—Data and observations. *Astron. Astrophys. Suppl.* **97**, 887–936 (1993).
5. Thornley, M. D. Uncovering spiral structure in flocculent galaxies. *Astrophys. J.* **469**, L45–L48 (1996).
6. Digel, S., de Geus, E. & Thaddeus, P. Molecular clouds in the extreme outer galaxy. *Astrophys. J.* **422**, 92–101 (1994).
7. Braine, J., Brouillet, N. & Baudry, A. The anatomy of an isolated spiral galaxy: NGC 4414. *Astron. Astrophys.* **318**, 19–28 (1997).
8. Cuillandre, J., Lequeux, J., Allen, R. J., Mellier, Y. & Bertin, E. Gas, dust, and young stars in the outer disk of M31. *Astrophys. J.* **554**, 190–201 (2001).
9. van Zee, L., Salzer, J. J., Haynes, M. P., O'Donoghue, A. A. & Balonek, T. J. Spectroscopy of outlying H II regions in spiral galaxies: abundances and radial gradients. *Astron. J.* **116**, 2805–2833 (1998).
10. Engargiola, G., Plambeck, R. L., Rosolowsky, E. & Blitz, L. Giant molecular clouds in M33. I. BIMA all-disk survey. *Astrophys. J. Suppl.* **149**, 343–363 (2003).
11. Heyer, M. H. & Terebey, S. The anatomy of the Perseus spiral arm: 12CO IRAS imaging observations of the W3–W4–W5 cloud complex. *Astrophys. J.* **502**, 265–277 (1998).
12. Blitz, L., Spergel, D. N., Teuben, P. J., Hartmann, D. & Burton, W. B. High-velocity clouds: building blocks of the local group. *Astrophys. J.* **514**, 818–843 (1999).
13. Dickman, R. L., Snell, R. L. & Schloerb, F. P. Carbon monoxide as an extragalactic mass tracer. *Astrophys. J.* **309**, 326–330 (1986).
14. Sodroski, T. J. *et al.* A three-dimensional decomposition of the infrared emission from dust in the Milky Way. *Astrophys. J.* **480**, 173–187 (1997).
15. Spitzer, L. *Physical Processes in the Interstellar Medium* 333 (Wiley-Interscience, New York, 1978).
16. Dickey, J. M. & Brinks, E. An H I absorption survey of M31 and M33—The mixture of warm and cool phases in the interstellar media of spiral galaxies. *Astrophys. J.* **405**, 153–162 (1993).
17. Braun, R. The temperature and opacity of atomic hydrogen in spiral galaxies. *Astrophys. J.* **484**, 637–655 (1997).
18. Pogge, R. W. Ionized gas in the nuclear regions of nearby non-Seyfert spiral galaxies. *Astrophys. J. Suppl.* **71**, 433–453 (1989).
19. Braine, J., Combes, F. & van Driel, W. NGC 4414: A flocculent galaxy with a high gas surface density. *Astron. Astrophys.* **280**, 451–467 (1993).
20. Pfenniger, D. & Combes, F. Is dark matter in spiral galaxies cold gas? Fractal models and star non-formation. *Astron. Astrophys.* **285**, 94–118 (1994).

Acknowledgements We thank J.-C. Cuillandre for taking the CFHT image and T. Osterloo, G. Gentile and G. Jozsa for making the H I data available. This work is based on observations carried out with the IRAM 30-m telescope. IRAM is supported by INSU/CNRS (France), MPG (Germany), and IGN (Spain).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence or requests for materials should be addressed to J.B. (braine@obs.u-bordeaux1.fr).

Electrical generation and absorption of phonons in carbon nanotubes

B. J. LeRoy, S. G. Lemay, J. Kong & C. Dekker

Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, 2628 CJ, Delft, The Netherlands

The interplay between discrete vibrational and electronic degrees of freedom directly influences the chemical and physical properties of molecular systems. This coupling is typically studied through optical methods such as fluorescence, absorption and Raman spectroscopy. Molecular electronic devices provide new opportunities for exploring vibration–electronic interactions at the single molecule level^{1–6}. For example, electrons injected from a scanning tunnelling microscope tip into a metal can excite vibrational excitations of a molecule situated in the gap between tip and metal⁷. Here we show how current directly injected into a freely suspended individual single-wall carbon nanotube can be

used to excite, detect and control a specific vibrational mode of the molecule. Electrons tunnelling inelastically into the nanotube cause a non-equilibrium occupation of the radial breathing mode, leading to both stimulated emission and absorption of phonons by successive electron tunnelling events. We exploit this effect to measure a phonon lifetime of the order of 10 ns, corresponding to a quality factor of well over 10,000 for this nanomechanical oscillator.

Single-wall carbon nanotubes (SWCNTs) were grown by chemical vapour deposition on a Pt substrate, which had predefined 100-nm-wide trenches etched in it. The SWCNTs were measured as-grown without any further processing. Details of the preparation of the samples used in this experiment have been reported previously⁸. Figure 1a sketches the set-up for the measurements indicating the scanning tunnelling microscope (STM) tip and the SWCNT crossing a trench. Figure 1b shows an STM topographic image of a nanotube suspended across a trench. The SWCNT is suspended over a distance of about 100 nm. Figure 1c is a zoom-in on the tube in the region over the trench, demonstrating atomic resolution.

An STM can reveal information about electronic structure with high spatial resolution through tunnelling spectroscopy. Figure 2a plots the normalized tunnelling differential conductance, $(dI/dV)/(I/V)$, as a function of sample voltage near the edge of a trench for the semiconducting SWCNT of Fig. 1b. As the sample voltage is decreased, a series of sharp spikes are obtained owing to the Coulomb staircase^{8,9}. The spacing between peaks is a measure of the energy necessary to add an electron to the SWCNT. Figure 2b plots the normalized differential conductance at the centre of the trench. New side peaks appear in addition to the previously observed Coulomb peaks. Such additional peaks in the differential conductance can occur when there are new channels through which electrons can tunnel onto the SWCNT. Figure 2c shows the full spatial dependence of the tunnelling differential conductance. The four main Coulomb staircase peaks shift in energy owing to a changing tip–SWCNT capacitance⁸ but persist throughout the image. In contrast, the additional side peaks are localized in the part of the SWCNT that is suspended. That side peaks only occur in the suspended portion explains why they have not been observed in previous STM measurements where nanotubes are lying on a conducting surface.

An STM allows the resistance of one of the tunnel barriers to be changed, thus controlling the current through the SWCNT at a fixed

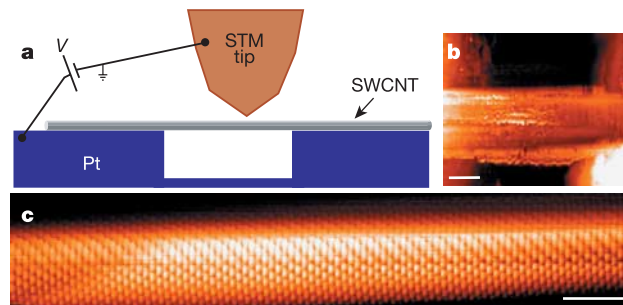


Figure 1 Measurement set-up and topographic images. **a**, Schematic diagram showing the set-up for performing spectroscopy on suspended SWCNTs. A voltage is applied to the substrate with respect to the tip, and the current flowing from the substrate through the SWCNT to the tip is measured. **b**, STM image of a nanotube crossing a trench. Scale bar, 25 nm. The apparent width of the 2-nm-diameter tube is enlarged by tip convolution. **c**, High-resolution image of the suspended portion of the SWCNT showing atomic resolution. Scale bar, 2 nm. The STM images were taken with a feedback current of 300 pA at -1 V. All of the measurements were performed at 5 K in an ultrahigh-vacuum STM.

applied voltage. On the basis of previous transport measurements, the resistance of the SWCNT–substrate barrier is known to be in the range 10–1,000 k Ω . Since the resistance of the tip–SWCNT barrier (~ 1 G Ω) is much greater than the SWCNT–substrate barrier, the total resistance and tunnel rate is set by the former barrier. Figure 3a plots the normalized differential conductance through a metallic SWCNT as a function of bias voltage for a low setpoint current, $I_{\text{set}} = 100$ pA, and a sample voltage of -0.6 V. The peaks due to the Coulomb staircase appear in groups of four, indicating that there are two spin-degenerate bands in this metallic SWCNT (refs 10 and 11, and S. Sapmaz, manuscript in preparation). Figure 3b and c plots the differential conductance with a medium setpoint current of 300 pA and a high current of 1,000 pA, respectively. Many additional peaks have appeared on either side of the main Coulomb staircase peaks. The peaks are equally spaced in energy on both sides of the main Coulomb peaks. Although the side-peak intensities change, the energy relative to the Coulomb peak is found to be independent of the current.

We attribute these side peaks to phonon-assisted tunnelling into the SWCNT. By absorbing or emitting a phonon, electrons tunnelling onto the SWCNT can increase or decrease their energy by $\hbar\omega$, where ω is the frequency of the phonon. Side peaks thus appear at energies $\hbar\omega$ from the main Coulomb peaks. Figure 3d is a zoom-in on one of the Coulomb peaks showing side peaks corresponding to both emission and absorption of a phonon, in direct analogy to the well-known Stokes and anti-Stokes peaks in Raman spectroscopy. The main Coulomb peak occurs when electrons tunnel elastically from the STM tip to the SWCNT (Fig. 3f). Peaks at energies above the Coulomb peaks (farther from $V = 0$) are due to electrons emitting a phonon when tunnelling (Fig. 3e). Likewise, side peaks on the low-energy side of the Coulomb peaks (closer to $V = 0$) correspond to electrons absorbing a phonon from the SWCNT when tunnelling (Fig. 3g). Note that the observation of a strong

peak for phonon absorption is very surprising given the fact that at equilibrium, without current, the population of phonons is extremely small at the low temperatures (5 K) of our experiment. The appearance of such peaks implies that current through the SWCNT induces a non-equilibrium phonon distribution. The phonon absorption effect has not previously been observed in electrical transport measurements. The non-equilibrium phonon distribution is furthermore responsible for the enhancement of the peaks on the high-energy side through stimulated emission.

It is possible to identify the specific phonon mode that we study in this experiment. For the 2.5-nm-diameter SWCNT of Fig. 2, the energy is found to be $E = 11.8 \pm 1.4$ meV, which corresponds to the

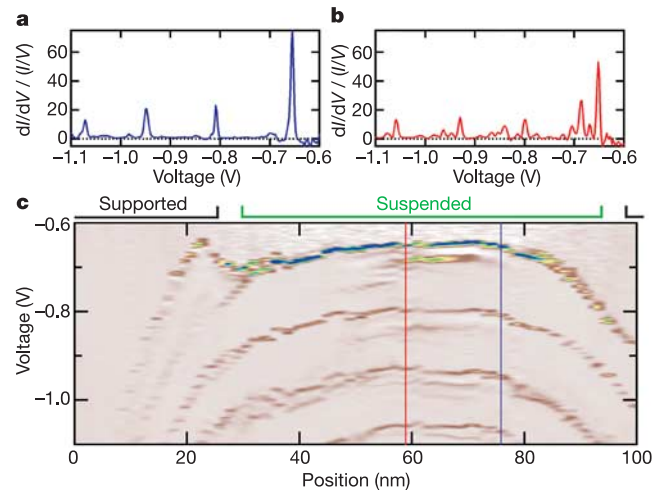


Figure 2 Spatially resolved spectroscopy along the suspended semiconducting SWCNT shown in Fig. 1b. **a**, Spectroscopy along the blue line in **c**, showing four spikes associated with the Coulomb staircase. The Coulomb staircase behaviour is determined by the two capacitances and resistances; the tip–SWCNT, and the SWCNT–substrate. The peaks occur when the Fermi level of the substrate aligns with states in the conduction band of the SWCNT. **b**, Spectroscopy along the red line in **c**, showing side peaks in addition to the Coulomb staircase peaks. **c**, Plot of the normalized differential conductance (colour scale) as a function of voltage and position. Sharp spikes are visible at all positions due to the Coulomb staircase, while extra peaks are visible only in the centre of the suspended region. The differential conductance was measured using lock-in detection with a 2 mV r.m.s. excitation voltage. The setpoint current was 300 pA at -1.25 V. The coloured lines above **c** indicate the regions where the SWCNT is supported (black) and suspended (green).

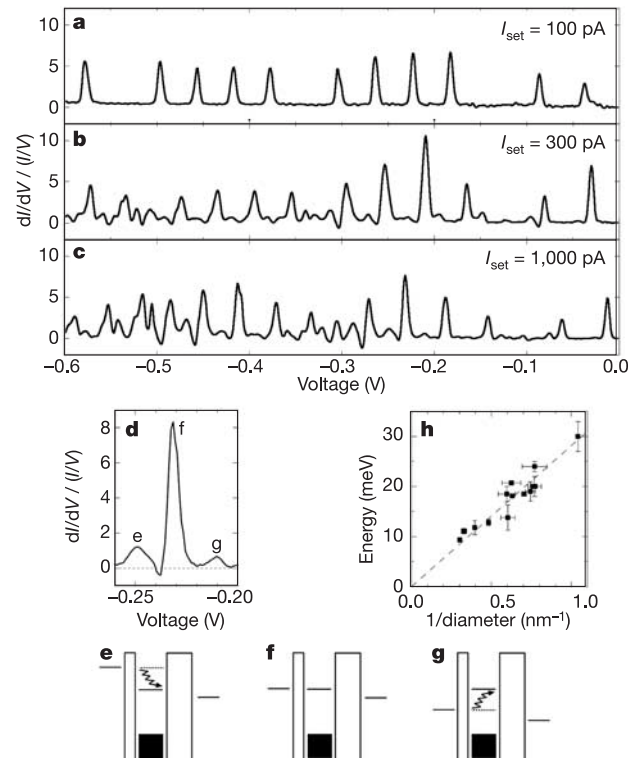


Figure 3 Current and diameter dependence of phonon-assisted tunnelling. **a**, Normalized differential conductance of a metallic SWCNT as a function of sample voltage taken with a low setpoint current $I_{\text{set}} = 100$ pA at -0.6 V. The tip is located at the centre of the suspended SWCNT. A series of sharp peaks is visible due to the Coulomb staircase as the Fermi level of the substrate aligns with unoccupied states of the SWCNT. **b, c**, Same as **a** with increasing setpoint current. A series of side peaks has appeared near the main Coulomb staircase peaks due to absorption and emission of phonons. To convert from sample voltage to energy, the capacitances between the tip and SWCNT, and the SWCNT and substrate must be known. The voltage dropped across the substrate–SWCNT junction is the applied voltage times the ratio of the tip–SWCNT capacitance and the total capacitance. These capacitances were determined from the spacing of the Coulomb peaks and the conductance between peaks^{8,19}. **d**, Zoom-in on one of the peaks showing side peaks corresponding to the emission (E) and absorption (G) of phonons. The data show a negative differential resistance between the main peak and the emission peak, which is often observed in these experiments. **e**, Energy diagram for emission of a single phonon showing that an increased energy is needed for electron tunnelling. The distance between the solid black line and the top of the black box represents the Coulomb charging energy. **f**, Energy diagram for elastic tunnelling where a level in the SWCNT is aligned with the leads. **g**, Energy diagram for absorption of a phonon, which decreases the energy needed for tunnelling. **h**, Plot of the energy of the side peaks as a function of inverse diameter, showing a linear relationship. The horizontal error bars arise from the width of the van Hove singularities, while the vertical error bars are the standard deviation of the side-peak energy. The dashed line is the expected energy for the radial breathing mode obtained from ref. 13, and is plotted without any adjustable parameters.

energy of the radial breathing mode (RBM). To verify the origin of the side peaks as the RBM phonon, we have repeated these measurements for a series of metallic and semiconducting SWCNTs. The measured energy of the side peaks as a function of inverse SWCNT diameter is plotted in Fig. 3h. The diameter of the SWCNTs is determined from the spacing of the first van Hove singularities¹². The dashed line is the theoretical energy dependence of the RBM, $27.8/d$ meV, where d is the diameter of the SWCNT in nm (ref. 13). The measured data points are best fitted with the equation $28.1 \pm 0.8 \text{ meV}/d$, which agrees very well with theory as well as with previously obtained results using Raman spectroscopy¹³. This demonstrates that the dominant phonon mode excited by electrons tunnelling into the SWCNT is the RBM, in agreement with other recent observations¹⁴. Electrons that tunnel into the SWCNT move at the Fermi velocity, thus travelling 10 nm in ~ 10 fs while the period of the RBM oscillation is ~ 370 fs. This suggests that circumferentially symmetric and low- k values of phonons are preferentially excited. This explains why the RBM is observed while the many other modes in the phonon density of states are not. When considering the excitation of low- k phonons and bending modes, we can exclude acoustic phonons as these will not be observed because of their low energy.

The new peaks that are observed for increased values of the current allow us to estimate the lifetime of the phonon in an individual SWCNT. The fact that side peaks on the low-energy

side are visible only for faster rates of electron tunnelling implies that the decay time for the phonon mode in this SWCNT is longer than the average time between electron tunnelling events at these rates. For example, for the case of Fig. 3b, the current for the first absorption peak visible at -0.15 V gives an average rate of electron tunnelling events of $1.44 \times 10^8 \text{ s}^{-1}$. This gives a decay time of at least $\tau \approx 7$ ns, which is more than 50 times longer than observed in Raman experiments on bundles of SWCNTs¹⁵. The decay time in the Raman data may be caused by coupling to other SWCNTs or the substrate which are absent in our case. Our measured decay time corresponds to a lower bound for the quality factor $Q = \tau E/h$ of 20,000 for the RBM. Typical values of Q that we obtain for the RBM in our devices are in the range 5,000–30,000. These large Q values are an order of magnitude larger than previous reports for other phonon modes in SWCNTs¹⁶, and are of the same order as those found in lithographically fabricated nanomechanical oscillators¹⁷.

To further quantify the effect of the tunnelling current on the vibrational excitations of the SWCNT, we have compared the measured differential conductance as a function of current with a simple model of phonon-assisted tunnelling. Figure 4a shows a specific differential-conductance peak as a function of voltage for a series of different currents. As the current is increased, the strength of the main Coulomb peak decreases whereas the peak associated with phonon emission increases. This shows that the probability of emitting a phonon is controlled by the rate of electrons passing through the SWCNT. The amplitude of these two peaks as a function of current is plotted in Fig. 4b, showing the side peak becoming stronger than the main peak. Figure 4c and d plots the differential conductance as a function of voltage and current for a different SWCNT. As the current is increased, an increasing number of side peaks (up to 4) appears. This shows that multi-phonon excitations become possible at high current levels.

In analogy to photon-assisted tunnelling, we can model the effect of the phonons as an oscillating potential on the SWCNT proportional to $\alpha \cos(\omega t)$, where α is the strength of the perturbation and ω is the frequency of the phonon mode¹⁸. This model predicts that the Coulomb peak at energy ϵ_0 evolves into a series of peaks $\sum_n J_n^2(\alpha) \frac{\partial f}{\partial V}(eV - \epsilon_0 - n\hbar\omega)$, where f is the Fermi function, J_n is the n th order Bessel function of the first kind, and n labels the peak number. The electron-phonon coupling term in the hamiltonian is linear in the phonon annihilation and creation operators, and therefore the strength of the perturbation $\alpha \propto \sqrt{N}$, where N is the number of phonons. Because the tunnel current I excites the phonons, a simple estimate is that $N \propto I$ and therefore $\alpha = \sqrt{I/\gamma}$, where γ is an unknown fitting parameter. The dashed lines in Fig. 4b show the expected peak height for the main peak and the first excited state as a function of current using this model. This simple model captures the essential trends in the experimental results. A better quantitative agreement to the experimental results can be obtained, however, if we assume that $N \propto I^2$. This gives a perturbation that is linear in the current, $\alpha = I/\beta$, where β is an unknown fitting parameter. The results of the simulations using this relationship for α are shown by the solid lines in Fig. 4b and e. An excellent agreement is obtained with only two largely uncorrelated fitting parameters, the peak height at zero current and β . As the current is increased in both the experimental data and the simulations, the number and strength of side peaks associated with phonon-assisted tunnelling increase, indicating that it is possible to control the population of phonons excited in a SWCNT.

Summing up, we have demonstrated that electrons passing through an individual carbon nanotube can populate phonon modes of the nanotube. This distribution can be probed with tunnelling spectroscopy that shows that electrons can both absorb and emit phonons when tunnelling onto the SWCNT. More theoretical work is needed to understand the current dependence of the phonon-assisted tunnelling peaks. These new phenomena

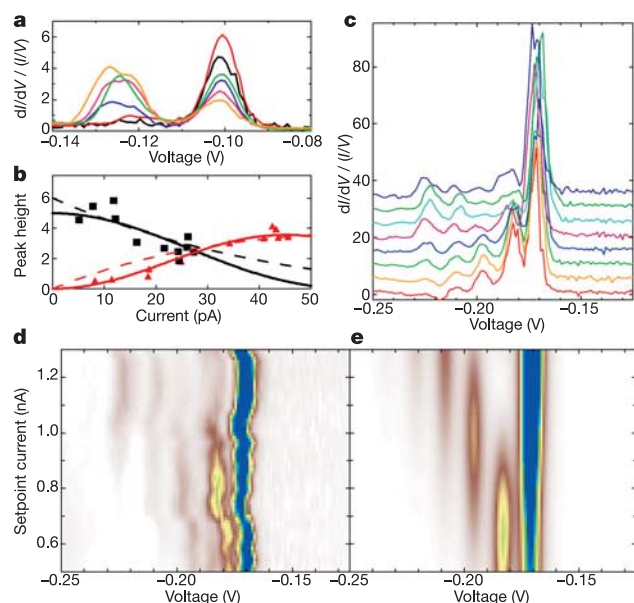


Figure 4 Comparison of the observed current dependence with theory. **a**, Normalized differential conductance versus bias voltage for a series of different currents ranging from 5 to 50 pA. As the current is increased, the main peak at -0.10 V decreases while the side peak at -0.125 V, associated with emission of a phonon, increases. **b**, Strength of the main (black squares) and first side peak (red triangles) as a function of the current at the given peak along with a fit using Bessel functions as described in the text. Dashed lines are for the perturbation $\alpha = \sqrt{I/\gamma}$ with $\gamma = 20$ pA, while solid lines are for $\alpha = I/\beta$ with $\beta = 24.5$ pA. **c**, Differential conductance as a function of bias voltage for a series of increasing currents from 500 to 1,300 pA, for a different SWCNT to **a**. **d**, Experimentally measured dependence of the differential conductance on current showing additional side peaks appearing as the current is increased. **e**, Simulation of the effect of increasing the current using $\alpha = I/\beta$ with $\beta = 310$ pA, which changes the phonon population and allows multiple phonon excitations. The conductance for the n th peak is given by the square of the n th-order Bessel function of the first kind. The data for **d** and **e** concern the first peak of a semiconducting SWCNT with zero current in the gap, and therefore there are no side peaks associated with phonon absorption (that is, on the right side of the main peak).

allow the measurement of long phonon lifetimes and high quality factors. Furthermore, it opens the door to the design of experiments with a known and controlled phonon population. □

Received 7 June; accepted 20 September 2004; doi:10.1038/nature03046.

1. Park, H. *et al.* Nanomechanical oscillations in a single- C_{60} transistor. *Nature* **407**, 57–60 (2000).
2. Pasupathy, A. N. *et al.* Vibration-assisted electron tunneling in C_{140} single-molecule transistors. Preprint at (<http://arxiv.org/cond-mat/0311150>) (2003).
3. Park, J. *et al.* Coulomb blockade and the Kondo effect in single-atom transistors. *Nature* **417**, 722–725 (2002).
4. Zhitenev, N. B., Meng, H. & Bao, Z. Conductance of small molecular junctions. *Phys. Rev. Lett.* **88**, 226801 (2002).
5. Qiu, X. H., Nazin, G. V. & Ho, W. Vibronic states in single molecule electron transport. *Phys. Rev. Lett.* **92**, 206102 (2004).
6. Weig, E. M. *et al.* Single-electron-phonon interaction in a suspended quantum dot phonon cavity. *Phys. Rev. Lett.* **92**, 046804 (2004).
7. Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. *Science* **280**, 1732–1735 (1998).
8. LeRoy, B. J., Lemay, S. G., Kong, J. & Dekker, C. Scanning tunneling spectroscopy of suspended single-wall carbon nanotubes. *Appl. Phys. Lett.* **84**, 4280–4282 (2004).
9. Grabert, H. & Devoret, M. H. *Single Charge Tunneling* (Plenum, New York, 1992).
10. Liang, W. J., Bockrath, M. & Park, H. Shell filling and exchange coupling in metallic single-walled carbon nanotubes. *Phys. Rev. Lett.* **88**, 126801 (2002).
11. Cobden, D. H. & Nygard, J. Shell filling in closed single-wall carbon nanotube quantum dots. *Phys. Rev. Lett.* **89**, 046803 (2002).
12. Venema, L. C., Meunier, V., Lambin, Ph. & Dekker, C. Atomic structure of carbon nanotubes from scanning tunneling microscopy. *Phys. Rev. B* **61**, 2991–2996 (2000).
13. Dresselhaus, M. S. & Eklund, P. C. Phonons in carbon nanotubes. *Adv. Phys.* **49**, 705–814 (2000).
14. Vitali, L. *et al.* Phonon spectromicroscopy of carbon nanostructures with atomic resolution. *Phys. Rev. Lett.* **93**, 136103 (2004).
15. Iliev, M. N., Litvinchuk, A. P., Arepalli, S., Nikolaev, P. & Scott, C. D. Fine structure of the low-frequency Raman phonon bands of single-wall carbon nanotubes. *Chem. Phys. Lett.* **316**, 217–221 (2000).
16. Reulet, B. *et al.* Acoustoelectric effects in carbon nanotubes. *Phys. Rev. Lett.* **85**, 2829–2832 (2000).
17. Cleland, A. *Foundations of Nanomechanics* (Springer, Berlin, 2003).
18. Tien, P. K. & Gordon, J. P. Multiphoton process observed in the interaction of microwave fields with the tunneling between superconductor films. *Phys. Rev.* **129**, 647–651 (1963).
19. Hanna, A. E. & Tinkham, M. Variation of the Coulomb staircase in a two-junction system by fractional electron charge. *Phys. Rev. B* **44**, 5919–5922 (1991).

Acknowledgements We thank Ya. M. Blanter for discussions, and NWO and FOM for funding.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.D. (dekker@mb.tn.tudelft.nl).

Random quasi-phase-matching in bulk polycrystalline isotropic nonlinear materials

M. Baudrier-Raybaut¹, R. Haïdar¹, Ph. Kupecsek^{1,2}, Ph. Lemasson³ & E. Rosencher^{1,4}

¹DMPH and DOTA/ONERA, Office National d'Etudes et de Recherches Aéropatiales, Chemin de la Hunière, 91761 Palaiseau, France

²Université Pierre et Marie Curie, 5 Place Jussieu, Paris 75005, France

³LPSC/CNRS, 1 Place Aristide Briand, F-92195, France

⁴Département de Physique, Ecole Polytechnique, 91228 Palaiseau, France

Three-wave mixing in nonlinear materials—the interaction of two light waves to produce a third—is a convenient way of generating new optical frequencies from common laser sources. However, the resulting optical conversion yield is generally poor, because the relative phases of the three interacting waves change continuously as they propagate through the material¹. This phenomenon, known as phase mismatch, is a consequence of optical dispersion (wave velocity is frequency dependent), and is responsible for the poor optical conversion potential of isotropic nonlinear materials². Here we show that exploiting the random

motion of the relative phases in highly transparent polycrystalline materials can be an effective strategy for achieving efficient phase matching in isotropic materials. Distinctive features of this 'random quasi-phase-matching' approach are a linear dependence of the conversion yield with sample thickness (predicted in ref. 3), the absence of the need for either preferential materials orientation or specific polarization selection rules, and the existence of a wavelength-dependent resonant size for the polycrystalline grains.

Much effort has recently been devoted to the development of materials which are suitable for nonlinear optical frequency conversion from the near- to the mid-infrared regions^{4–7}. Many semiconductor (GaAs, ZnSe, ...) materials of space group $\bar{4}3m$ are excellent candidates: they are widespread and mature optoelectronics graded materials, they are transparent in the mid-infrared and they display particularly high nonlinear 2nd-order susceptibilities (that is, for three-wave interaction). In single-crystalline materials, the main factor for an efficient optical conversion is the coherence length Λ_c , the distance over which the relative phase lag of the three waves add up to π . Indeed, because of the different relative phase velocities between the three interacting waves, optical power flows back and forth from the converted waves towards the pumping waves (that is, backconversion) as soon as the interacting distance becomes larger than the coherence length (see Fig. 1). Phase-matching is thus obtained when the coherence length is much longer than the interaction distance in the material. Because of the lack of optical birefringence, such a situation cannot be obtained naturally in isotropic $\bar{4}3m$ semiconductors over long distance. Instead, so-called quasi-phase-matching scenarios based on epitaxial growth on patterned substrates need to be developed^{1,8–10}.

The backconversion process results from an interference effect between the three coherent waves. Such interference could be destroyed if the waves were allowed to lose their respective phases randomly in the material, in that the nonlinear susceptibility does not average to zero. Such parametric interactions in disordered media have attracted considerable attention in recent years, with a particular emphasis on the nonlinear diffusion/scattering processes^{11–14}. In these studies, the average size of the nanocrystal-

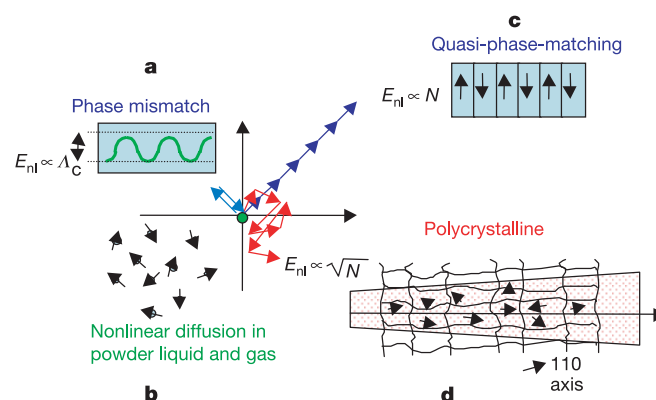


Figure 1 Three-wave mixing mechanisms in bulk, powder, periodically poled and polycrystalline materials. In bulk nonlinear optical materials, the relative phase of the three interacting fields grows continuously with the propagating distance, gaining a phase lag of π every coherence length Λ_c . The transfer of energy between the waves thus oscillates with a period of $2\Lambda_c$, leading to a small (if not null) conversion efficiency (a). In quasi-phase-matched materials¹⁰, the orientation of the crystal is rotated every Λ_c , compensating the phase lag of π so that the energy transfer E_{nl} adds up constructively with the propagating distance ($E_{nl} \propto N$, where N is the number of grains) (c). In a totally disordered material (powder, gas, liquid), each particle behaves independently, scattering the nonlinearly generated fields in an incoherent way (b). In polycrystalline materials, the relative phase diffuses in phase space, leading to a coherent growth of the nonlinear generated fields according to $E_{nl} \propto \sqrt{N}$.

lites is smaller than the interacting wavelengths (Rayleigh regime) so that the samples are highly diffusive and mostly opaque. We show that a similar (but different) strategy can be implemented in polycrystalline ZnSe materials, for which highly transparent materials and controllable grain size are available. The concept of random quasi-phase-matching is experimentally demonstrated in a difference frequency generation (DFG) experiment—a particularly interesting situation which involves large coherence lengths—with emphasis on the influence of the grain size, light polarization and length of the sample, suggesting the potential for new optical random materials¹⁵. We thus show that a signal far higher than the contribution of a single coherence length can be obtained, as a result of phase randomization due to the random distribution of the microcrystallite ('grain') domains³ (see Fig. 1).

We first theoretically describe the DFG process in polycrystalline materials. Let E_1^n be the value of the electromagnetic field at the output of the n th grain crossed by the two undepleted laser beams E_2 and E_3 . By integration over each grain of size X_m (with a gaussian distribution around its mean value) and summation of all the contributions, as already suggested by refs 3 and 13, the value of E_1^n is deduced from the classical growing term², taking into account randomizing phase-distribution terms:

$$E_1^n = \frac{\omega_1}{n_1 c} E_3 E_2^* \sum_{m=1}^n d_m \frac{e^{-i\Delta k X_m} - 1}{\Delta k} e^{-i\Delta k \sum_{j=1}^{m-1} X_j} \quad (1)$$

where c is the velocity of light, E_1, E_2, E_3 designate the DFG and the pump-wave fields of respective angular frequencies ω_1, ω_2 and ω_3 , n_1, n_2, n_3 are the optical indexes of the polycrystalline materials for the three waves, $\Delta k = k_3 - k_1 - k_2$ is the phase mismatch wave-vector between the three waves in each grain and d_m is the nonlinear coefficient of the m th grain. Numerical simulations show that the term

$$e^{-i\Delta k \sum_{j=1}^{m-1} X_j}$$

in equation (1) is such that, after a few grains, the three waves display a random relative phase while entering each micro-crystal of random size X . This is a key element for the success of random quasi-phase-matching. Performing an ensemble average of the square modulus of equation (1)—meaning that the beam waist is far larger than the mean grain size—the total DFG intensity

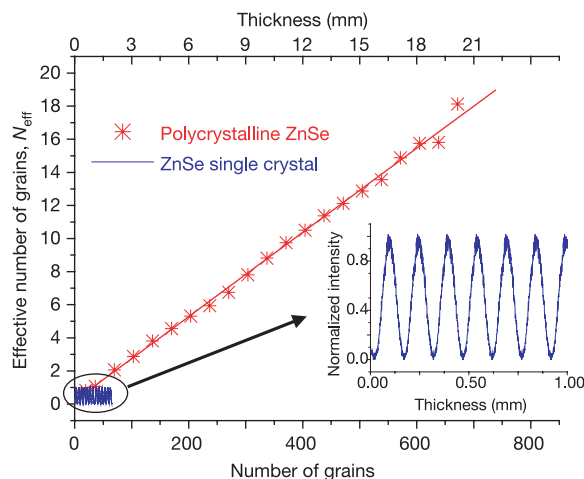


Figure 2 Variation of the normalized DFG intensity I_1 as a function of the polycrystalline sample thickness. The pump wavelengths are 1.925 and 2.380 μm , respectively, yielding a 10- μm generated field. The distance is expressed in terms of the number of ZnSe grains along the optical path ($\Lambda = 30 \mu\text{m}$ in this sample). The DFG signal is normalized to the intensity obtained with a single coherence length in a ZnSe single crystal. The variation of the DFG signal in ZnSe single crystalline materials is shown for comparison. The specificity of random quasi-phase-matching appears clearly.

generated by N grains is then given by:

$$I_1^{\text{poly}} = \frac{8\pi^2}{\lambda_1^2} Z_0 \frac{\langle |d|^2 \rangle}{n_1 n_2 n_3} \langle X^2 \text{sinc}^2(\Delta k X/2) \rangle N I_2 I_3 \quad (2)$$

where Z_0 is the vacuum impedance (377 Ω) and λ_1 is the wavelength of field E_1 . The $\langle x \rangle$ notation designates the mean value of random variable x averaged over its probability distribution $p(x)$. For instance, $\langle |d|^2 \rangle$ is the norm of the effective nonlinear coefficient $\langle |d|^2 \rangle = \int |d(\theta, \varphi)|^2 p(\theta, \varphi) d\theta d\varphi$ averaged over all the possible orientations of the grains (θ and φ are the eulerian coordinates of the grain). More specifically, we made the assumption that the grain size obeys a gaussian distribution, that is, $p(X) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(X-\Lambda)^2/2\sigma_x^2}$ where Λ and σ_x are the mean value and the standard deviation of the grain size X respectively. Equation (2) can be written as:

$$I_1^{\text{poly}} = N_{\text{eff}} I_1^{\text{coh}} \quad (3)$$

where $I_1^{\text{coh}} = \frac{32}{\lambda_1^2} Z_0 \frac{|d|^2}{n_1 n_2 n_3} \Lambda_c^2 I_2 I_3$ is the DFG intensity generated by a single coherence length of the materials for the specific three-wave combination, Λ_c is the coherence length and N_{eff} is the effective number of grains participating to the DFG process, given by:

$$N_{\text{eff}} = N \frac{\langle |d|^2 \rangle}{d^2} \langle \sin^2(\Delta k X/2) \rangle = N \frac{\langle |d|^2 \rangle}{d^2} \frac{1}{\sqrt{2\pi}\sigma_x} \int_0^\infty \sin^2(\Delta k X/2) e^{-(X-\Lambda)^2/2\sigma_x^2} dX \quad (4)$$

Equations (2)–(4) are very predictive. First, as already predicted in somewhat similar situations^{3,13,16}, a linear variation of the DFG signal as a function of the sample thickness L (where $L = N\Lambda$) is expected, instead of a quadratic variation for perfect phase matching¹. We call this regime random quasi-phase-matching (see Fig. 1). Second, the DFG signal never averages to zero whatever the grain orientation distribution. Finally, a careful examination of equation (4) shows that a resonance of the DFG yield is expected once the condition $\Delta k \Lambda = \pi$ is obtained, that is, when the average grain size is close to the DFG coherence length.

The starting polycrystalline ZnSe materials were provided by II–VI, Inc. It consists of micrograins with an average size of $\Lambda = 30 \mu\text{m}$, and a mean standard deviation $\sigma_x \approx 5\% \Lambda$. No important crystalline texture is revealed by X-ray Bragg diffraction studies, so that the $\langle |d|^2 \rangle/d^2$ averages to 0.14 (d is the ZnSe nonlinear susceptibility in the $\langle 110 \rangle$ direction and for all the polarization configurations of the interactive waves¹⁷). Samples were submitted to a solid phase recrystallization¹⁸ process. Four different mean grain sizes have thus been obtained: 30 μm (unprocessed), 60 μm , 70 μm and 100 μm . The grain size X is also gaussian distributed,

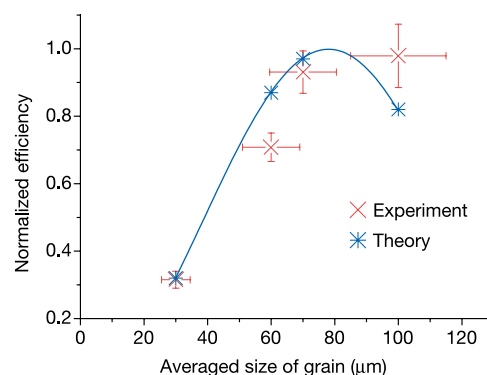


Figure 3 Normalized difference frequency generation efficiency as a function of the ZnSe mean grain size Λ . The different samples are obtained by solid phase recrystallization (straight line, theory; crosses and bars, experimental data points with error bars). A resonant yield appears when the grain size is close to the coherence length (78 μm , in this experimental scheme). Horizontal error bars indicate the standard deviation of the grain size; vertical error bars indicate the standard error of measurement of our experiment.

with a uniform standard deviation of $5\ \text{\AA}/100$. To study the DFG conversion efficiency as a function of the sample thickness L , the samples are bevelled by mechanical polishing to produce wedges with angles of 25° for the recrystallized samples and 37° for the unprocessed ones.

For the DFG experiment, the pump source is a 100-mJ, 30-Hz, 1.06- μm pumped LiNbO₃ type I optical parametric oscillator with signal and idler waves tunable between 1.8 and 2.4 μm , yielding an 8–12- μm DFG field. See ref. 19, where a similar set-up has been used to measure DFG coherence length in ZnSe ($\sim 78\ \mu\text{m}$ in the 8–12- μm DFG range)¹⁹. If a resonance is to be found, it is thus expected for a grain size of about 78 μm . This large value of the coherence length originates from the minimum index dispersion of ZnSe in this spectral region equidistant between the gap and the Reststrahlung energies. This explains the large nonlinear yields expected in ZnSe polycrystalline materials, as expected from equations (1) to (4). About 200 μJ of the available energy are focused to a waist of about 200 μm in the samples. The bevelled samples are translated on a motorized mount. The DFG wave energy is measured using a HgCdTe cryogenic detector protected by adequate filters.

We first studied the variation of the DFG signal as a function of the sample thickness. To explore a large thickness range, we chose the 37° wedged sample. Figure 2 shows the variation of the measured signal as a function of the crystal thickness. The signal was normalized to the signal obtained in the same configuration by a DFG signal due to a signal coherence length (as determined in a crystalline $\langle 110 \rangle$ wedge sample¹⁹). Clearly, the signal follows the expected linear dependence on sample length L . Given the experimental values $\Lambda/\Lambda_c \approx 0.38$, $\sigma_x/\Lambda \approx 5\%$ and $\langle d^2 \rangle/d^2 \approx 14\%$, equation (4) predicts a N_{eff}/N ratio of 4.5%, whereas a ratio of 2.8% is experimentally determined from Fig. 2, which is the right order of magnitude.

We then studied the influence of the grain size on the DFG signal. The preceding experiment was performed for each of the three SPR samples. Linear variations as a function of the sample thickness were observed in all the samples. Figure 3 shows the theoretical and experimental values of the normalized efficiency for the four different values of the grain mean size Λ , taking into account an experimentally determined loss term of $\rho = 4.5\% \text{ cm}^{-1}$. This loss is due to a very small amount of light scattering (corresponding to a mean free path for $\sim 2\text{-}\mu\text{m}$ photons of 22 cm), a situation which is fundamentally different from previous work¹⁴. The observed variations are in good agreement with the theoretical predictions. Finally, we study the variation of the DFG signal as a function (1) of the orientation of the pump polarization relative to the sample bevel direction, (2) of the angle between the pump and the DFG beam polarization, and (3) of the two pump frequencies. No significant variation of the DFG signal could be observed in any of these experiments, which confirms the absence of crystallographic texture in our samples.

Random quasi-phase-matching may then become a valuable building block in nonlinear optics. Such materials can be deposited on any substrate (such as silicon) with no particular restrictions regarding crystalline growth or sample length, opening the way to optical conversion in low-cost deposited waveguides. Its extremely loose frequency selectivity makes it of particular interest for generating optical radiations with ultra-wide spectral tunability and reasonable efficiency. It may also be applied to emerging materials such as ZnO, chromium-doped ZnSe or new laser ceramics, allowing multifunctional materials to be developed.

We note that, although quasi-phase-matched GaAs crystals have been demonstrated to be far more efficient nonlinear optical converters in the 8–12 μm range^{8,9,20}, these crystals have not yet been demonstrated over large lengths. But large size samples are available in polycrystalline materials (commonly $> 100\ \text{mm}$ for the ZnSe samples), so that good conversion yields could eventually be reached. This, added to the extreme ease-of-use of the random

quasi-phase-matching technique (almost no control is needed), makes of any piece of polycrystalline ZnSe a cheap and efficient optical converter. $\square$

Received 3 May; accepted 22 September 2004; doi:10.1038/nature03027.

1. Armstrong, J. A., Bloembergen, N., Ducuing, J. & Pershan, P. S. Interactions between light waves in a nonlinear dielectric. *Phys. Rev.* **127**, 1918–1939 (1962).
2. Rosencher, E. & Vinter, B. *Optoelectronics* (Cambridge Univ. Press, Cambridge, 2002).
3. Morozov, E. Y., Kaminskii, A. A., Chirkin, A. S. & Yusupov, D. B. Second optical harmonic generation in non linear crystals with a disordered domain structure. *JETP Lett.* **73**, 647–650 (2001).
4. Fejer, M. M. Nonlinear optical frequency conversion. *Phys. Today* **40**, 25–32 (1994).
5. Ebrahimzadeh, M. & Dunn, M. H. Parametric generation of tunable light from continuous-wave to femtosecond pulses. *Science* **286**, 1513–1517 (1999).
6. Fiore, A., Berger, V., Rosencher, E., Bravetti, P. & Nagle, J. Phase matching using an isotropic nonlinear optical material. *Nature* **391**, 463–466 (1998).
7. Rosencher, E. *et al.* Quantum engineering of optical nonlinearities. *Science* **271**, 168–173 (1996).
8. Levi, O. *et al.* Difference frequency generation of 8- μm radiation in orientation-patterned GaAs. *Opt. Lett.* **27**, 2091–2093 (2002).
9. Eyres, L. A. *et al.* All-epitaxial fabrication of thick, orientation-patterned GaAs films for nonlinear optical frequency conversion. *Appl. Phys. Lett.* **79** (2001).
10. Fejer, M. M., Magel, G. A., Jundt, D. H. & Byer, R. L. Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE J. Quant. Electron.* **28**, 2631–2654 (1992).
11. Agranovich, V. M. & Kravtsov, V. E. Nonlinear backscattering from opaque media. *Phys. Rev. B* **43**, 13691–13694 (1991).
12. Kravtsov, V. E., Agranovich, V. M. & Grigorishin, K. I. Theory of second-harmonic generation in strongly scattering media. *Phys. Rev. B* **44**, 4931–4942 (1991).
13. Makeev, E. V. & Skipetrov, S. E. Second harmonic generation in suspensions of spherical particles. *Opt. Commun.* **224**, 139–147 (2003).
14. Mel'nikov, V. A. *et al.* Second-harmonic generation in strongly scattering porous gallium phosphide. *Appl. Phys. B* **79**, 225–228 (2004).
15. Wiersma, D. S. & Cavalieri, S. A temperature-tunable random laser. *Nature* **414**, 708–709 (2002).
16. Kurtz, S. K. & Perry, T. T. A powder technique for the evaluation of non linear optical materials. *J. Appl. Phys.* **39**, 3798–3813 (1968).
17. Shoji, I., Kondo, T., Kitamoto, A., Shirane, M. & Ito, R. Absolute scale of second-order nonlinear-optical coefficients. *J. Opt. Soc. Am. B* **14**, 2268–2294 (1997).
18. Rzepka, E., Roger, J. P., Lemasson, P. & Triboulet, R. Optical transmission of ZnSe crystals grown by solid phase recrystallisation. *J. Cryst. Growth* **197**, 480–484 (1999).
19. Haidar, R. *et al.* Largely tunable mid-infrared (8–12 μm) difference frequency generation in isotropic semiconductors. *J. Appl. Phys.* **91**, 2550–2552 (2002).
20. Vodopyanov, K. L. *et al.* Optical parametric oscillation in quasi-phase-matched GaAs. *Opt. Lett.* **29**, 1912–1914 (2004).

Acknowledgements We are indebted to C. Sanchez and A. Cheniere for X-ray measurements, A. Godard, M. Lefebvre and N. Guérineau for help, to M. Fejer for discussions, and D. Sessler for critical reading of the manuscript. This work was supported by the Délégation Générale pour l'Armement (DGA).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E. R. (rosencher@onera.fr).

Metal wires for terahertz wave guiding

Kanglin Wang & Daniel M. Mittleman

Department of Electrical and Computer Engineering, MS 366, Rice University, Houston, Texas 77251-1892, USA

Sources and systems for far-infrared or terahertz ($1\ \text{THz} = 10^{12}\ \text{Hz}$) radiation have received extensive attention in recent years, with applications in sensing, imaging and spectroscopy^{1–10}. Terahertz radiation bridges the gap between the microwave and optical regimes, and offers significant scientific and technological potential in many fields. However, waveguiding in this intermediate spectral region still remains a challenge. Neither conventional metal waveguides for microwave radiation, nor dielectric fibres for visible and near-infrared radiation can be used to guide terahertz waves over a long distance, owing to the high loss from the finite conductivity of metals or the high

absorption coefficient of dielectric materials in this spectral range. Furthermore, the extensive use of broadband pulses in the terahertz regime imposes an additional constraint of low dispersion, which is necessary for compatibility with spectroscopic applications. Here we show how a simple waveguide, namely a bare metal wire, can be used to transport terahertz pulses with virtually no dispersion, low attenuation, and with remarkable structural simplicity. As an example of this new waveguiding structure, we demonstrate an endoscope for terahertz pulses.

In an attempt to meet the compelling need for useful THz waveguides, various guides with quasi-optical coupling have been demonstrated within the last few years. Most of these have been based on conventional guiding structures, such as metal tubes^{11,12}, plastic ribbons¹³ or dielectric fibres¹⁴. There have also been reports on the application of the latest technology of photonic crystal fibres to THz radiation^{15,16}. In all of these cases, the utility for transport of THz pulses is limited not only by the high loss, but also by group velocity dispersion of the guided waves. The most promising studies have reported dispersionless propagation in parallel metal plate

waveguides^{17–19}. But in this case the attenuation is still unacceptably high, due in large part to the finite conductivity of the metal plates. In addition, the cross-sectional area of the waveguide is too large for many of the proposed THz applications, including in particular medical diagnostics.

Our interest in the wire geometry was motivated by our recent studies of the propagation of THz pulses on optical antennas in near-field scanning optical microscopy (NSOM)²⁰. Here, with the direct measurement of the electric field in the time domain, we show that the guided mode on a metal wire is similar to the transverse electromagnetic (TEM) mode of a conventional coaxial waveguide, and exhibits all of the properties required for a practical far-infrared waveguide. Because of the low exposed metal surface area of these wire waveguides, the attenuation due to conductivity losses are substantially lower than any previously reported structure. Based on these unique properties, we have realized a Y-splitter and a 90° output directing structure, and have constructed the first endoscope for THz pulses.

In our experiment, the broadband single-cycle pulses of free-space THz radiation are generated and coherently detected using ultrafast photoconductive sampling^{2,3} with fibre-coupled THz antennas²¹. A diagram of the experimental set-up is shown in Fig. 1. The horizontally polarized THz pulses are focused onto the stainless steel waveguide. A second stainless steel wire is placed at the focal spot, oriented perpendicular to the waveguide (that is, out of the page in Fig. 1). This second wire serves as an input coupler. Scattering of the input THz radiation at the intersection structure helps to excite a propagating waveguide mode, with a radially polarized mode pattern. Both the waveguide and the coupler are 0.9 mm in diameter, and the separation between them is 0.5 mm. The receiver is placed at the end of the waveguide, and is oriented to detect only the vertically polarized component of the electric field in order to eliminate the possibility of detecting directly scattered radiation which would interfere with the detection of the guided mode. The incident THz beam is modulated by a chopper in front of the transmitter, and a lock-in amplifier is used for detecting the induced photocurrent in the receiver. The THz transmitter, the focusing lenses, and the coupler are all mounted on a movable stage so that the incident position along the waveguide can be controlled. The THz receiver is mounted on a three-axis stage for detection at various positions relative to the end of the waveguide.

Figure 2a shows typical time-domain electric field waveforms, for two different receiver positions located symmetrically above and below the wire waveguide. These waves are vertically (y) polarized, perpendicular to the horizontally (x) polarized input beam. The polarity reversal as the detector scans across the wire clearly shows the radially polarized nature of the guided wave. To observe the spatial profile of the mode, we scan the receiver in a plane perpendicular to the waveguide axis, with a time delay fixed at the

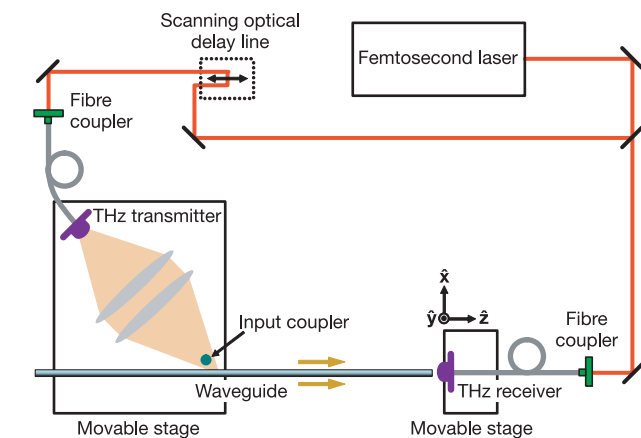


Figure 1 A diagram of the optical set-up for characterizing the propagating electromagnetic mode on a metal wire waveguide. Horizontally polarized THz pulses are generated by a fibre-coupled photoconductive transmitter and focused onto a stainless steel waveguide with a diameter of 0.9 mm. A second stainless steel wire is placed at the focal spot, oriented perpendicular to the waveguide, to act as an input coupler. A radially polarized mode is excited in the space around the waveguide. The electric field of the propagating pulses is detected at the end of the waveguide with a fibre-coupled photoconductive receiver, which is sensitive only to the vertical polarization component. The transmitter, the focusing lenses, and the coupler are all mounted on a movable stage which can be moved along the waveguide. The receiver is mounted on a stage which can be moved in three dimensions relative to the end of the waveguide.

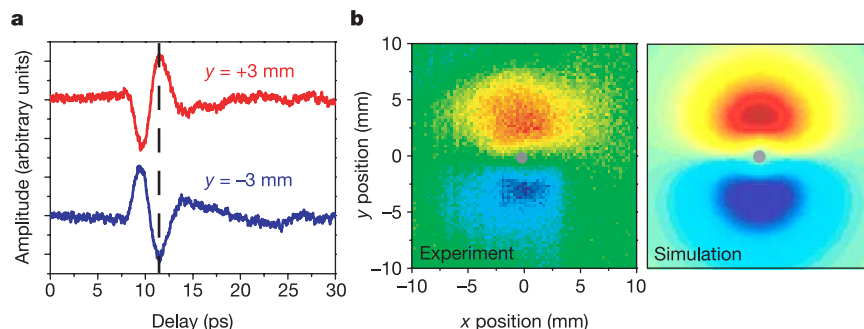


Figure 2 Spatial mode of the guided wave on a metal wire. **a**, Time-domain electric field waveforms detected with the receiver 3 mm above and 3 mm below the waveguide. The polarity reversal is an indication of the radial nature of the polarization. **b**, Left panel: spatial profile of the electric field obtained by moving the THz receiver in a plane

perpendicular to the waveguide axis. The time delay is fixed at the peak of the THz pulses indicated by the dashed line in **a**. Red represents positive values and blue represents negative values. The simulation (right panel) shows the vertical component of the electric field for a cylindrically symmetric radial mode.

peak of the THz pulses. Because the detector is sensitive only to one polarization, it is not possible to directly detect the spatial distribution of a radially polarized mode. Instead, the measured mode should resemble the projection of a cylindrically symmetric radial mode onto the vertical (y) direction. We simulate this effect by first computing the two-dimensional field distribution of a radial mode and then projecting it onto a single polarization axis. This simulation is in good agreement with the data (see Fig. 2b and also ref. 22).

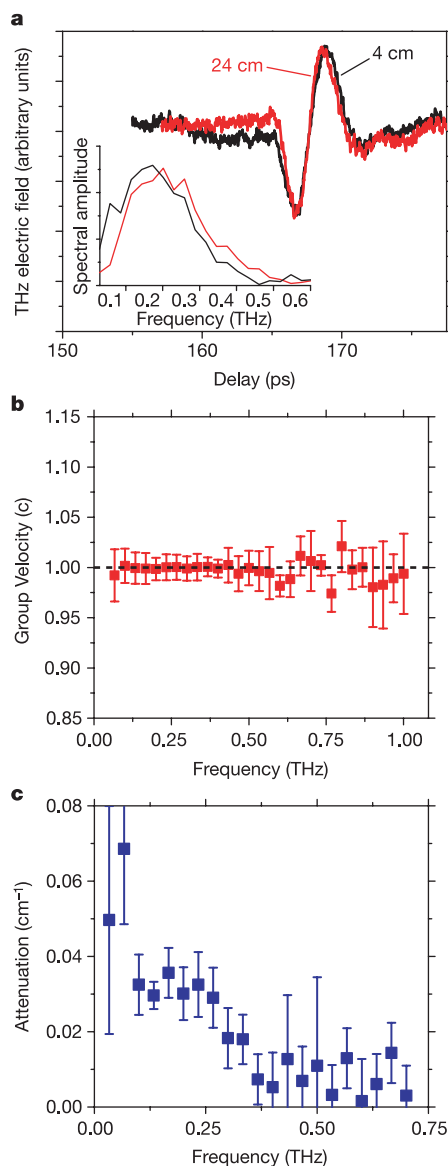


Figure 3 Characteristics of the propagating mode. **a**, THz waveforms measured after 4 cm (black) and 24 cm (red) of propagation distance along the wire. In order to compare the shapes of these two waveforms, the red curve has been shifted by 667.9 ps along the time axis, so that it coincides with the black curve. In addition, it has been multiplied by a factor of 1.6, which corresponds to an amplitude attenuation coefficient of less than 0.03 cm⁻¹. The propagation is essentially dispersionless. The red curve is slightly narrower, indicating that the high frequency components are attenuated slightly less than the low frequency components. This is also clear from the normalized amplitude spectra of the two waveforms, shown in the inset. **b**, Group velocity of the propagating mode as a function of frequency, derived from a series of measurements with different propagation distances. **c**, The electric field amplitude attenuation coefficient of the propagating mode as a function of frequency, derived in the same manner as in **b**. As anticipated from **a**, this loss decreases with increasing frequency, in contrast to other THz waveguides in which the loss is dominated by finite conductivity effects. In **b** and **c**, error bars show ± 1 s.d.

We characterize the propagation of this radial mode by moving the incident position of the THz beam along the waveguide. There is no evident change in the shape of the time-domain waveforms for propagation distances up to 24 cm, showing that this propagation is largely dispersionless (see Fig. 3a). By analysing the spectra of these waveforms, we can derive the group velocity, v_g , and the electric field amplitude attenuation coefficient, α , as a function of frequency. These results are shown in Fig. 3b and c. The radial nature of the mode, combined with the absence of group velocity dispersion throughout the measurable spectral range, suggests that the propagating mode has the characteristics of the lowest-order (TEM) mode of a coaxial waveguide. The average attenuation coefficient, weighted by the pulse power spectrum, is less than 0.03 cm⁻¹, the lowest of any THz waveguide reported to date¹⁹.

This low attenuation emphasizes one unique aspect of the wire waveguide. Compared to other waveguide geometries, a metal wire has a much smaller surface area interacting with the electromagnetic field, so the propagation loss due to finite conductivity of the metal is negligible²³. Instead, the measured radiative losses arise from diffractive spreading of the propagating mode in the lateral dimensions. This distinction can be seen by noting that, as shown in Fig. 3, the losses decrease with increasing frequency, in contrast to the case of waveguides in which the loss is dominated by ohmic effects¹⁸. This frequency dependence may arise from the overlap between the guided mode and propagating far-field modes. We note that our measurements do not reflect the losses associated with the coupling of the free-space THz beam to the guided mode; in the experiment described here, these losses are quite large. In addition, our coupling mechanism probably excites a superposition of guided modes, which may lead to larger diffractive losses due to the higher transverse spatial frequencies that comprise higher-order coaxial modes. More effective mode-matching will be important to optimize both the input coupling and the propagation losses^{12,24}.

We observe that the guided mode can propagate on a slightly curved waveguide without a substantial increase in the loss. It is also found that the guided propagation can be easily coupled between two curved waveguides in contact with each other (or between a curved waveguide and a straight one). These features enable a very simple scheme for a beam splitter, as illustrated in Fig. 4. Another remarkable feature of the metal wire waveguide is that the mode maintains its radial nature as it propagates off the end of the

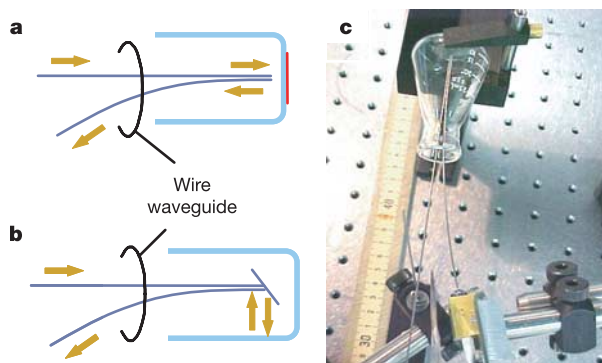


Figure 4 The THz endoscope. **a**, The optical configuration for a THz endoscopic measurement of a region at the bottom of a cavity. The input pulses are launched to the detection region with a straight waveguide. By using a Y-splitter structure, the reflected pulses are directed to a branch waveguide for detection. The red bar indicates a metal plate affixed to the exterior of the cavity, to enhance the reflection from this dielectric interface. **b**, The optical configuration for a THz endoscopic measurement of the side wall in a cavity. The input pulses are launched by a straight waveguide, and directed off the waveguide perpendicularly by a small mirror attached at the distal end of the endoscope. The reflected pulses go back to the waveguide after the reflection at the distal mirror, and are then directed to the output waveguide by a Y-splitter. **c**, Photograph of a THz endoscope which is inserted into a flask, for a measurement with the configuration shown in **a**.

waveguide, for at least several centimetres. As a result, we can build a 90° output director by attaching a small mirror at the end of the waveguide, tilted at a 45° angle with respect to the axis of the waveguide. With these two structures realized, we construct a THz endoscope based on the wire waveguide. Endoscopic measurements can be performed by inserting the distal end of the endoscope (with attached mirror) into any container. Two such configurations are shown in Fig. 4.

Figure 5 shows the results of demonstration experiments performed in these two configurations. In each case, we display a pair of waveforms measured on either side of the waveguide, as in Fig. 2a. In Fig. 5a, the two reflections in each waveform arise from the two dielectric interfaces at the bottom of the glass flask. By attaching a piece of metal to the flask bottom, the strength of the second reflection is enhanced. Another experiment (Fig. 5b) is the measurement of a reflection from the side wall of a metal tube into which the endoscope has been inserted, as in Fig. 4b. In this case, the signal is not as strong as in the previous measurement owing to the additional propagation and coupling processes.

The low loss and negligible group velocity dispersion make this new type of waveguide especially suitable for use in THz research and applications. Our results also show that it is now possible to direct THz pulses inside containers or around corners, where line-of-sight optical access is not practical. In addition, we can launch THz pulses along any thin metal rod structures. In situations where the guided mode could be perturbed by other structures close to the waveguide, a section of outer metallic shield could be added to form a coaxial waveguide, as long as the additional ohmic losses can be tolerated. A THz endoscope has been successfully constructed to

demonstrate the use of the metal wire waveguide, which paves the way for a wide range of new applications for terahertz sensing and imaging. □

Received 13 July; accepted 13 September 2004; doi:10.1038/nature03040.

1. Mittleman, D. (ed.) *Sensing with Terahertz Radiation* (Springer, Heidelberg, 2002).
2. Smith, P. R., Auston, D. H. & Nuss, M. C. Subpicosecond photoconducting dipole antennas. *IEEE J. Quant. Electron.* **24**, 255–260 (1988).
3. van Exter, M. & Grischkowsky, D. Characterization of an optoelectronic terahertz beam system. *IEEE Trans. Microwave Theory Tech.* **38**, 1684–1691 (1990).
4. Jepsen, P. U., Jacobsen, R. H. & Keiding, S. R. Generation and detection of terahertz pulses from biased semiconductor antennas. *J. Opt. Soc. Am. B* **13**, 2424–2436 (1996).
5. Mittleman, D. M., Jacobsen, R. H. & Nuss, M. C. T-ray imaging. *IEEE J. Select. Top. Quant. Electron.* **2**, 679–692 (1996).
6. Jacobsen, R. H., Mittleman, D. M. & Nuss, M. C. Chemical recognition of gases and gas mixtures with terahertz waves. *Opt. Lett.* **21**, 2011–2013 (1996).
7. Woodward, R. M., Wallace, V. P., Arnone, D. D., Linfield, E. H. & Pepper, M. Terahertz pulsed imaging of skin cancer in the time and frequency domain. *J. Biol. Phys.* **29**, 257–261 (2003).
8. Crawley, D. et al. Three-dimensional terahertz pulse imaging of dental tissue. *J. Biomed. Opt.* **8**, 303–307 (2003).
9. Kawase, K., Ogawa, Y. & Watanabe, Y. Non-destructive terahertz imaging of illicit drugs using spectral fingerprints. *Opt. Express* **11**, 2549–2554 (2003).
10. Wang, S. & Zhang, X.-C. Pulsed terahertz tomography. *J. Phys. D* **37**, R1–R36 (2004).
11. McGowan, R. W., Gallot, G. & Grischkowsky, D. Propagation of ultrawideband short pulses of THz radiation through submillimeter-diameter circular waveguides. *Opt. Lett.* **24**, 1431–1433 (1999).
12. Gallot, G., Jamison, S. P., McGowan, R. W. & Grischkowsky, D. Terahertz waveguides. *J. Opt. Soc. Am. B* **17**, 851–863 (2000).
13. Mendis, R. & Grischkowsky, D. Plastic ribbon THz waveguides. *J. Appl. Phys.* **88**, 4449–4451 (2000).
14. Jamison, S. P., McGowan, R. W. & Grischkowsky, D. Single-mode waveguide propagation and reshaping of sub-ps terahertz pulses in sapphire fiber. *Appl. Phys. Lett.* **76**, 1987–1989 (2000).
15. Han, H., Park, H., Cho, M. & Kim, J. Terahertz pulse propagation in a plastic photonic crystal fiber. *Appl. Phys. Lett.* **80**, 2634–2636 (2002).
16. Goto, M., Quema, A., Takahashi, H., Ono, S. & Sarukura, N. Teflon photonic crystal fiber as terahertz waveguide. *Jpn. J. Appl. Phys.* **43**, L317–L319 (2004).
17. Mendis, R. & Grischkowsky, D. Undistorted guided-wave propagation of subpicosecond terahertz pulses. *Opt. Lett.* **26**, 846–848 (2001).
18. Mendis, R. & Grischkowsky, D. THz interconnect with low loss and low group velocity dispersion. *IEEE Microwave Wireless Comp. Lett.* **11**, 444–446 (2001).
19. Coleman, S. & Grischkowsky, D. A THz transverse electromagnetic mode two-dimensional interconnect layer incorporating quasi-optics. *Appl. Phys. Lett.* **83**, 3656–3658 (2003).
20. Wang, K., Barkan, A. & Mittleman, D. M. Propagation effects in apertureless near-field optical antennas. *Appl. Phys. Lett.* **84**, 305–307 (2004).
21. Rudd, J. V., Zimdars, D. & Warmuth, M. Compact fiber-pigtailed terahertz imaging system. *Proc. SPIE* **3934**, 27–35 (2000).
22. Quabis, S., Dorn, R., Eberler, M., Glöckl, O. & Leuchs, G. Focusing light to a tighter spot. *Opt. Commun.* **179**, 1–6 (2000).
23. Marcuvitz, N. *Waveguide Handbook* (McGraw-Hill, New York, 1951).
24. Coleman, S. & Grischkowsky, D. Parallel plate THz transmitter. *Appl. Phys. Lett.* **84**, 654–656 (2004).

Acknowledgements This work was supported in part by the R. A. Welch Foundation, the National Science Foundation and Advanced Micro Devices.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.M.M. (daniel@rice.edu).

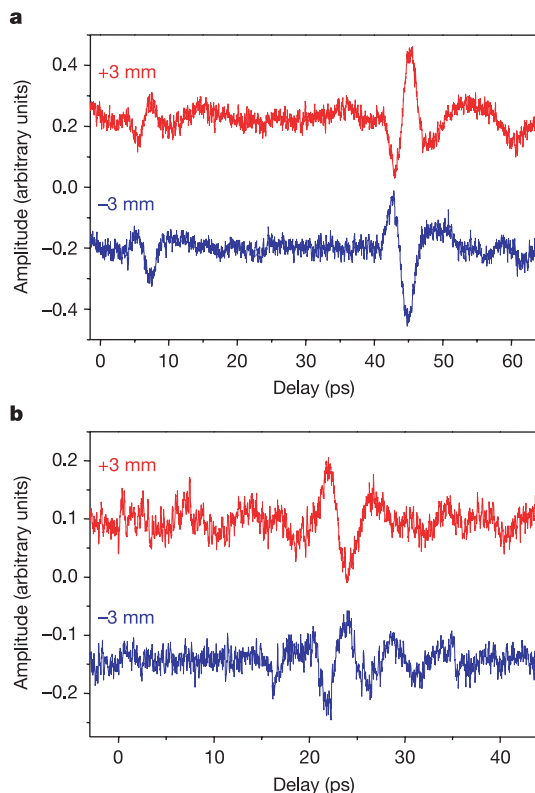


Figure 5 Endoscopic measurements. **a**, THz waveforms obtained with the receiver 3 mm above and 3 mm below the end of the output waveguide, in the experiment depicted in Fig. 4c. The two pulses in each waveform arise from the reflections at the two dielectric interfaces at the bottom of the glass flask. The enhancement of the second pulse is due to a metal plate attached to the flask bottom. **b**, THz waveforms obtained in a measurement with the configuration of Fig. 4b, by inserting the endoscope with an end directing mirror into a metal tube. The relatively weak signal results from the additional propagation and coupling process of THz pulses at the distal end of the endoscope.

Constraints on the duration and freshwater release of Heinrich event 4 through isotope modelling

D. Roche, D. Paillard & E. Cortijo

IPSL/Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS, 91191 Gif sur Yvette, France

Heinrich events¹—abrupt climate cooling events due to ice-sheet instability that occurred during the last glacial period—are recorded in sediment cores throughout the North Atlantic Ocean^{2,3}. Modelling studies have described likely physical mechanisms^{4–6} for these events, but the quantitative characteristics of

Heinrich events are less well known. Here we use a climate model of intermediate complexity⁷ that explicitly calculates the distribution of oxygen isotopes in the oceans to simulate Heinrich event 4 at about 40,000 yr ago. We compare an ensemble of scenarios for this Heinrich event with oxygen isotope data measured in foraminiferal calcite of a comprehensive set of sediment cores^{8,9}. From this comparison, we obtain a duration of 250 ± 150 yr and an ice release of 2 ± 1 m sea-level equivalent for Heinrich event 4, significantly reducing the uncertainties in both values compared to earlier estimates^{5,10–14} of up to 2,000 yr and 15 m of sea-level equivalent ice release, respectively. Our results indicate that the consequences of Heinrich events may have been less severe than previously assumed, at least with respect to Greenland climate and sea level.

Sediment cores from the North Atlantic covering the last glacial period show layers with high ice-rafted debris (IRD) content, known as the Heinrich layers¹. The debris has been recognized to be bedrock mainly from the Laurentide² ice sheet, but also from the Fennoscandian³ ice sheet, and was brought into the middle of the Atlantic by iceberg discharges. The melting of numerous icebergs in an area called the IRD belt (between 40°N and 55°N; ref. 15) not only produced IRD layers, but also provided a huge freshwater input to the North Atlantic. Modelling studies^{4,16,17} have shown that such inputs would cause the thermohaline circulation (THC) to collapse, considerably modifying climate over the North Atlantic^{18,19}. During Heinrich events, the $\delta^{18}\text{O}_c$, as measured in the calcite of polar planktonic foraminifera, shows a decrease of up to -2‰ in the area of the IRD belt⁸. This is consistent with the melting of icebergs, as ice sheets have a very low $\delta^{18}\text{O}$ content. In these data, there are no direct constraints on the volume of ice that melted in the Atlantic Ocean, and only a few constraints on the duration of such events. Assessments of the duration rely on different dating techniques, and yield an overall range between 50 and 2,000 yr (refs 5, 10–14). Those obtained with ^{14}C dating are longer (more than 1,000 yr) and uncertain owing to potential reservoir age problems during drastic ocean circulation changes such as Heinrich events²⁰. The threshold value for a THC shutdown in oceanic models is model dependent, ranging from 0.05 to 0.25 Sv (refs 4, 21, 22), and cannot be used directly to constrain the iceberg discharge rate. A simulation with a coupled climate–ice-sheet model²³ yields a duration of 100 to 1,000 yr and a contribution of about 5 m to sea-level rise. A more conceptual model of the cause of Heinrich events⁵ gives a duration of 250–550 yr and a released volume of $1.25 \times 10^6 \text{ km}^3$ as best estimates.

Figure 1 shows the data that we use here to characterize Heinrich event 4 (HE4). The $\delta^{18}\text{O}_c$ anomaly is maximum just north of 50°N, reaching 1.95‰. Southward, the anomaly decreases, with values of about 1‰ at around 45°N, and the anomaly drops to 0.2‰ around 41°N. Northward, the anomaly decreases to about 0‰ at 60°N.

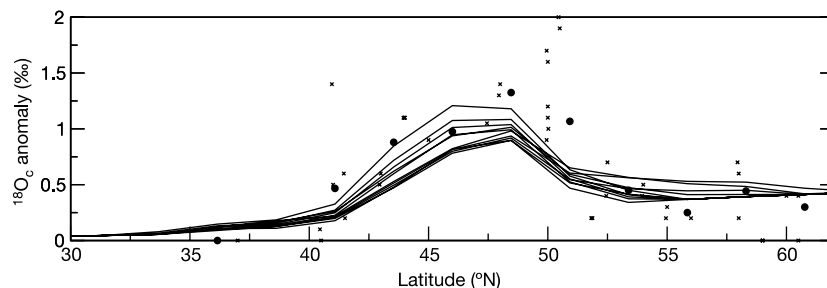


Figure 1 Data and modelled $\delta^{18}\text{O}_c$ anomaly as a function of latitude. Anomaly is shown as 'before the Heinrich event'—'during the Heinrich event'. The latter term refers to the local maximum of IRD in time; the former term is the last point before the IRD anomaly. Small symbols ('x') show the raw data^{8,9}; large symbols (filled circles) show the data averaged on

Further north, the signal shows possible influence from the Fennoscandian ice sheet, which we have not considered here. Here we focus on the 40–60°N latitudinal band. To compare this data set to our simulations, we averaged it on a 2.5° grid (see Fig. 1). Analysing HE4 with a zonally averaged point of view is legitimate as both the $\delta^{18}\text{O}_c$ anomaly and the estimated temperature anomalies are more or less zonally homogeneous (see Fig. 7A in ref. 8).

We used the CLIMBER-2 climate model of intermediate complexity⁷, in which we implemented the water isotopes²⁴, to compute the surface $\delta^{18}\text{O}_c$ during HE4. The isotopic content of water fluxes are computed in the model at the ocean surface consistently with the simulated climate. Classically⁴, we use a given freshwater input to represent the melting of icebergs. Little is known about the temporal dynamics of the freshwater input from ice sheets. Here, we model this input as a single pulse of prescribed cosine shape. To characterize it, two parameters are needed: the duration D_e and the maximum instantaneous flux M_f , the volume being the integral of the prescribed function (see Methods). The use of a different pulse shape gives similar results for comparable input parameters. As we do not know precisely the $\delta^{18}\text{O}_w$ of melt water, we assume a constant value of -30‰ . This value is consistent with the $\delta^{18}\text{O}$ of ice for this period measured in mid- to low-altitude ice cores (for example, Dye 3 and Renland²⁵), but also with estimations from modelling studies²⁶. The uncertainty associated with this choice is about $\pm 5\text{‰}$, which would lead to ± 0.3 m of equivalent sea-level in our final estimation. A large number of transient experiments were integrated with D_e between 100 and 2,000 yr, and M_f between 0.1 and 2 Sv (that is, with a total freshwater volume between 0.1×10^6 and $7 \times 10^6 \text{ km}^3$).

To compare model results with empirical data, we define the similarity S between simulated and empirical values for $\delta^{18}\text{O}_c$ (see Methods). We obtain a maximum in the computed similarity (Fig. 2) in the flux-duration space for $100 < D_e < 400$ yr and $0.24 \leq M_f \leq 0.34$ Sv. Using a 90% confidence level, we find that the freshwater flux for HE4 is consistent with $D_e = 250 \pm 150$ yr and $M_f = 0.29 \pm 0.05$ Sv. In other words, we evaluate the input to be 1.9 ± 1.1 m of equivalent sea-level and the simulated temperature anomaly to be 6°C at 45°N . The THC is off for about 150 yr in a simulated Heinrich event of 300 yr.

Maximization of the similarity between data and model completely determines our input parameters. For D_e , we point out that, during a simulation, the anomaly first appears in the melting zone (centred at about 45°N). If the THC is stopped—or drastically reduced—the anomaly spreads northward and southward with diffusion and wind-driven circulation. If the THC is still significantly active, then the anomaly spreads northward with the global circulation, but not significantly southward. This means that the latitudinal extent of the anomaly south of the IRD belt is linked to

the CLIMBER grid. Analytical errors associated with the raw data are $< 0.1\text{‰}$. Lines show our best fit runs, having a similarity of more than $S = 0.26$ with data computed after a gaussian model (see text).

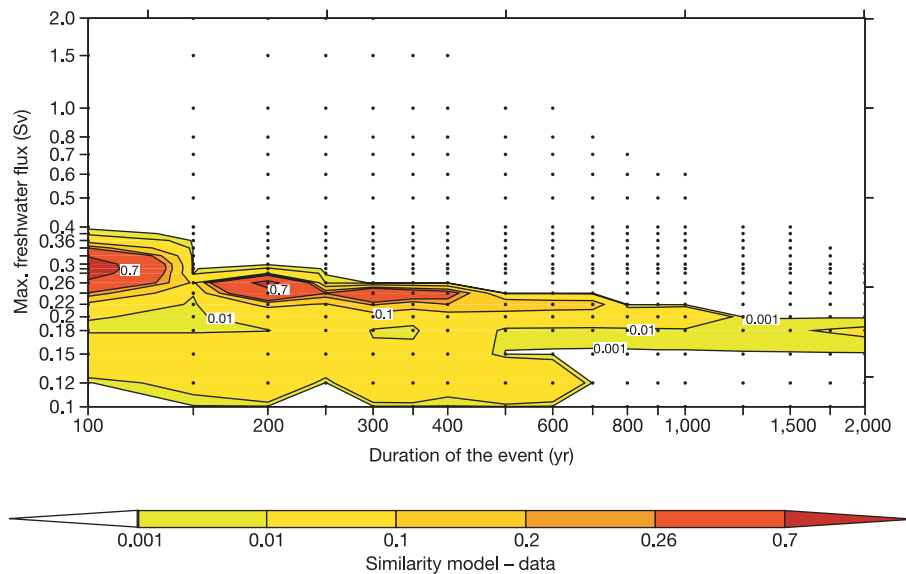


Figure 2 Computed similarity to data for each simulation plotted as a function of duration and maximum freshwater flux. The similarity is defined after a gaussian model (see Methods). Each dot represents one transient simulation. The upper-right corner of the figure is empty, as iceberg volume would have been greater than $7 \times 10^6 \text{ km}^3$. The best

fits are in red (highest similarity). The points with small similarity at lower fluxes (below the best fit) correspond to simulations for which the THC is not shut down. Points with small similarity at higher fluxes are points for which M_f is too high.

the duration for which the THC is off. If this duration is increased, then the southward extent of the anomaly will also be larger. In Fig. 2, all runs with THC still significantly active have a small similarity to data (for example, runs at $M_f = 0.1 \text{ Sv}$). This is consistent with data-based studies²⁷ which also indicate that the THC shuts down. When the maximum flux, M_f , is increased, then the anomaly simulated in the melting zone also increases; this is because the anomaly decays at a constant rate linked to diffusion and wind-driven circulation (with THC off). If M_f is too strong, the anomaly in this zone will be inconsistent with data. If diffusion and wind-driven circulation in the model are well constrained by comparing the model results to present-day measurements, the threshold for which the THC shuts down is a priori model dependent. This dependence is evaluated with two different approaches (see Methods).

Temperature is traditionally used as the major constraint for climate models, to be compared with temperature estimations from data. However, the simulations with CLIMBER-2 show that the temperature anomaly is only bimodal during a simulated Heinrich event: either the THC is off (or drastically reduced) and then we produce an anomaly broadly consistent with data (about 6°C at $40\text{--}45^\circ\text{N}$), or the THC is still active and there is no large temperature anomaly. This shows that comparing data and models in palaeoclimates is fruitful when tracers are simulated directly in models, whereas deriving model-like prognostic variables from data is more difficult.

In order to have the same duration as that estimated from ^{14}C dating, we considered some multi-meltwater-pulse experiments (each pulse with characteristics close to those determined above). To keep the simulated signal close to what we see in the data, the two events have to be separated by a restart of the THC, to ensure that the surface anomaly from the first meltwater pulse is subducted in the ocean before the incoming of the second one. In the model, we need up to 400 yr after the end of the first pulse to achieve this restart. So in 1,000 yr, we could potentially repeat twice a run similar to that performed above; that is, a contribution of about 4 m to sea-level rise. In very high resolution cores, with time resolution of the order of a century²⁸, no such bimodal signal can clearly be seen in $\delta^{18}\text{O}_c$. As the resolution of these cores is typically three times better

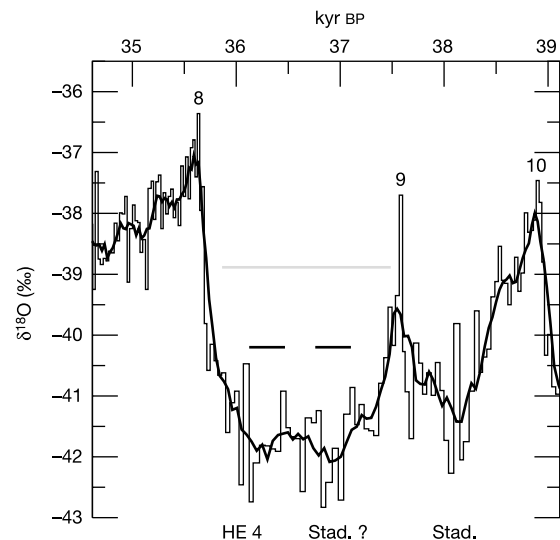


Figure 3 $\delta^{18}\text{O}$ from the GRIP ice-core¹⁹ record. The thin stepped black line shows the unaveraged data, whereas the thicker black line shows a 5-point running average of the same data. The thick grey bar indicates the 'classical' time interval given for Heinrich event 4¹⁸, whereas the two small thick black bars indicate HE4 as implied by our work. 'Stad.' shows position of stadials, 'HE4' the location of Heinrich event 4 as discussed in the text.

than the duration needed to restart the THC, we consider this scenario unlikely.

Our results are in good agreement with emerging sedimentological evidence^{13,14} that Heinrich events were shorter than a millennium. We have demonstrated that for HE4 consistency between model results and data can only be achieved by invoking a complete—or nearly complete—shutdown of the THC. These results emphasize that tracer simulation is needed for thorough analysis of palaeo-records, and this allows us to gain confidence in the model behaviour during rapid climate changes. There are broad

implications of these results for interpretation of the palaeo-record and for mechanisms of rapid climate change. One is shown in Fig. 3, where the 'classical' interval of 1,500 yr for HE4 between Dansgaard-Oeschger events 8 and 9 (see, for example, ref. 18) is compared to our new result. There are potentially two intervals of low ^{18}O lasting about three centuries during this period. Determining which one of these two intervals is HE4 would involve correlation of very high resolution oceanic cores with the ice-core record. In terms of mechanisms, a recent study²⁹ showed that a meltwater pulse from the Antarctic ice sheet could trigger a warming phase in the north, which could be the warming towards Dansgaard-Oeschger event 8. Such a meltwater discharge could have been triggered by a southern warming caused by a northern cooling (that is, via a see-saw mechanism) due to HE4. In this framework, we would then favour the low- ^{18}O period at 36.25 kyr before present to be the time of HE4 in the GRIP Greenland ice-core record. □

Methods

Freshwater forcing function

The freshwater flux is given to the model as the function FWF, defined as:

$$\text{FWF}(t) = \frac{1}{4} M_f \left(1 - \cos \left(2\pi \frac{t - T_{\max}}{D_e} \right) \right)^2$$

where time t is between $T_{\min}$, the beginning, and $T_{\max}$, the end, time of the perturbation. The duration, D_e , is then defined as $T_{\max} - T_{\min}$. The input volume is the time integral of this function over the perturbation interval, that is:

$$V(t) = \frac{3}{8} M_f D_e$$

For the latitude repartition of the anomaly, we kept the method of ref. 20, that is, a cosine of the latitude. However, in an attempt to be closer to the data, we have changed the location of the input freshwater flux, and centred it at 45°N.

Similarity between data and model run results

We first computed the variance of all available data points at each CLIMBER grid cell. For each grid cell, i , we average the available number, n_i , of data points. The variance associated with the scatter of the data at one CLIMBER grid point is then:

$$\sigma_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (d_k - \mu_i)^2$$

where the d_k are the data points in cell i , and μ_i the average of the data points in this cell. Then we compute the similarity following a gaussian model:

$$S = \exp \left[- \sum_{i=1}^N \frac{(s_i - \mu_i)^2}{2\sigma_i^2} \right]$$

where s_i , μ_i are the simulated and observed values of $\delta^{18}\text{O}_c$ in the model grid cell i respectively, and σ_i is the variance computed above. To compute S , we use the N model grid cells between 40 and 60°N ($N = 8$). The 90% confidence level following this gaussian model is at $S = 0.26$.

Evaluation of THC threshold sensitivity

To evaluate the dependence of our result on the threshold for which the THC shuts down, we have followed two different approaches. First, we have performed two additional sets of experiments where we added a constant additional freshwater flux of 0.1 and -0.1 Sv to the North Atlantic, thus modifying the sensitivity of the CLIMBER model to the freshwater forcing. The results obtained are almost identical to those obtained without this additional flux, suggesting the independence of the results to the THC shutdown threshold. Second, we have developed a simple analytical model of the surface North Atlantic, in which we can choose the THC shutdown threshold, and evaluate the effect on our results.

Our analytical model for ^{18}O is made up of a set of two linear equations, describing a system of two boxes. The northern box (box 1) represents the point of maximum anomaly (at about 47°N, see Fig. 1a), and the southern box (box 2) is at about 40°N. The two boxes exchange ^{18}O with an infinite underlying box of isotopic composition 0‰. Diffusion is considered between all the boxes. D_1 (D_2) is the diffusive flux between box 1 (box 2) and the infinite box. Diffusion (named K) also exists between boxes 1 and 2. An advection flux, called A , is also considered from the infinite box, through the southern and the northern boxes then back into the infinite box. This advection is used to have an idea of the role of THC circulation described in the text, and is fixed at zero or a non-zero value. Input of an anomalous flux of depleted isotopic composition ($\delta_{\text{ice}} = -30$ ‰) is given to the northern box to represent the Heinrich event as a constant value (input function is here stepwise). This two-box model leads to a system of ordinary differential equations, as follows:

$$d\delta_1(t)/dt = (A + K)\delta_2(t) - (A + K + D_1)\delta_1(t) + M_f\delta_{\text{ice}}/V$$

$$d\delta_2(t)/dt = -(A + K + D_1)\delta_2(t) + K\delta_1(t)$$

where $\delta_i(t)$ is the isotopic composition of box i , and V is the average volume of the two boxes. This system may be solved using the initial condition $\delta_1(0) = 0$ and $\delta_2(0) = 0$. The

isotopic anomaly produced by this simple two-box model is fitted to the results obtained with CLIMBER-2 in the same regions, in order to obtain close results. This fitting gives values for A , K , D_1 , D_2 and V . In this model, we can choose to shut down the THC (that is, set A to zero) and show the influence of this parameter on our results.

The results (not shown) indicate that an oceanic model with a threshold higher than about 0.35 Sv is not able to match the palaeodata (as the surface ^{18}O anomaly just after THC shutdown will already be too strong), whereas all models with smaller thresholds will have the same solution, which yield about 200 yr in the analytical model. Following this analytical model result, we can conclude that, on the order of a century, our estimation of HE4 is independent of the threshold for which the THC shuts down.

Received 9 April; accepted 15 September 2004; doi:10.1038/nature03059.

1. Heinrich, H. Origin and consequence of cyclic ice rafting in the northeast Atlantic Ocean during the past 130,000 years. *Quat. Res.* **29**, 142–152 (1988).
2. Broecker, W., Bond, G., Klas, M., Clark, E. & McManus, J. Origin of the northern Atlantic's Heinrich events. *Clim. Dyn.* **6**, 265–273 (1992).
3. Grousset, F. *et al.* Patterns of ice-rafted detritus in the glacial north Atlantic (40–55°N). *Paleoceanography* **8**, 175–192 (1993).
4. Ganopolski, A. & Rahmstorf, S. Rapid changes of glacial climate simulated in a coupled climate model. *Nature* **409**, 153–158 (2001).
5. MacAyeal, D. A low order model of the Heinrich Event cycle. *Paleoceanography* **8**, 767–773 (1993).
6. Paillard, D. & Labeyrie, L. Role of the thermohaline circulation in the abrupt warming after Heinrich Events. *Nature* **302**, 162–164 (1994).
7. Petoukhov, V. *et al.* CLIMBER-2: a climate system model of intermediate complexity. Part I: model description and performance for present climate. *Clim. Dyn.* **16**, 1–17 (2000).
8. Cortijo, E. *et al.* Changes in sea surface hydrology associated with Heinrich event 4 in the North Atlantic Ocean between 40° and 60°N. *Earth Planet. Sci. Lett.* **146**, 129–145 (1997).
9. Elliot, M., Labeyrie, L., Dokken, T. & Manthé, S. Coherent patterns of ice-rafted debris deposits in the Nordic regions during the last glacial (10–60 ka). *Earth Planet. Sci. Lett.* **194**, 151–163 (2001).
10. Bond, G. & Lott, R. Iceberg discharge into the north Atlantic on millennial time scale during the last glaciation. *Science* **267**, 1005–1010 (1995).
11. Elliot, M. *et al.* Millennial-scale iceberg discharges in the Irminger basin during the Last Glacial period: Relationship with the Heinrich events and environmental settings. *Paleoceanography* **13**, 433–446 (1998).
12. Grousset, F. Les changements abrupts du climat depuis 60 000 ans. *Quaternaire* **12**, 203–211 (2001).
13. Dowdeswell, J., Maslin, M., Andrews, J. & McCave, I. Iceberg production, debris rafting, and the extent and thickness of Heinrich layers (H-1, H-2) in North Atlantic sediments. *Geology* **23**, 301–304 (1995).
14. François, R. & Bacon, M. Heinrich events in the North Atlantic: radiochemical evidence. *Deep-Sea Res.* **41**, 315–334 (1994).
15. Ruddiman, W. Late quaternary deposition of ice-rafted sand in the sub-polar north Atlantic (lat 40° to 65°). *Geol. Soc. Am. Bull.* **88**, 1813–1821 (1977).
16. Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic Ocean. *Nature* **378**, 165–167 (1995).
17. Fanning, A. F. & Weaver, A. J. Temporal-geographical meltwater influences on the North Atlantic Conveyor: Implications for the Younger Dryas. *Paleoceanography* **12**, 307–320 (1997).
18. Bond, G. *et al.* Correlations between climate records from North Atlantic sediments. *Nature* **365**, 143–147 (1993).
19. Grootes, P., Stuiver, M., White, J., Johnsen, S. & Jouzel, J. Comparison of oxygen records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).
20. Stocker, T. & Wright, D. The effect of a succession of ocean ventilation changes on ^{14}C . *Radiocarbon* **40**, 359–366 (1998).
21. Seidov, D., Sarnthein, M., Statteger, K., Prien, R. & Weinelt, M. North Atlantic ocean circulation during the last glacial maximum and subsequent meltwater event: a numerical model. *J. Geophys. Res.* **101**, 16305–16332 (1996).
22. Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995).
23. Calov, R., Ganopolski, A., Petoukhov, V., Claussen, M. & Greve, R. Large-scale instabilities of the Laurentide ice sheet simulated in a fully coupled climate-system model. *Geophys. Res. Lett.* **29**, 2216 doi:10.1029/2002GL016078 (2002).
24. Roche, D., Paillard, D., Ganopolski, A. & Hoffmann, G. Oceanic oxygen-18 at the present-day and the LGM: equilibrium simulations with a coupled climate model of intermediate complexity. *Earth Planet. Sci. Lett.* **218**, 317–330 (2004).
25. Johnsen, S. *et al.* The $\delta^{18}\text{O}$ record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climate instability. *J. Geophys. Res.* **C 102**, 26397–26410 (1997).
26. Jouzel, J., Joster, R., Suozzo, R. & Russell, G. Stable water isotope behavior during the last glacial maximum: A general circulation model analysis. *J. Geophys. Res.* **D 99**, 25791–25801 (1994).
27. Vidal, L. *et al.* Evidence for changes in the North Atlantic Deep Water linked to meltwater surges during the Heinrich events. *Earth Planet. Sci. Lett.* **146**, 129–145 (1997).
28. Grousset, F. *et al.* Zooming in on Heinrich layers. *Paleoceanography* **16**, 240–259 (2001).
29. Weaver, A., Saenko, O., Clark, P. & Mitrovica, J. Meltwater pulse 1A from Antarctica as a trigger of the Bolling-Allerød warm interval. *Science* **299**, 1709–1713 (2003).

Acknowledgements We thank F. Jacquemart for comments on the mathematics of the box model; A. Ganopolski for technical assistance with respect to the CLIMBER-2 model; and M. Kagayama, G. Ramstein and Y. Donnadieu for comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.R. (roche@lsce.saclay.cea.fr).

Triassic marine reptiles gave birth to live young

Yen-nien Cheng¹, Xiao-chun Wu² & Qiang Ji³

¹National Museum of Natural Science, 1 Kuan Chien Road, Taichung 404, Taiwan, China

²Canadian Museum of Nature, P.O. Box 3443 STN 'D', Ottawa, Ontario, K1P 6P4, Canada

³Institute of Geology, Chinese Academy of Geological Sciences, 26 Baiwanzhuang Road, Beijing 100037, China

Sauropterygians form the largest and most diverse group of ancient marine reptiles that lived throughout nearly the entire Mesozoic era (from 250 to 65 million years ago)^{1,2}. Although thousands of specimens of this group have been collected around the world since the description of the first plesiosaur in 1821 (ref. 3), no direct evidence has been found to determine whether any sauropterygians came on shore to lay eggs (oviparity) like sea turtles, or gave birth in the water to live young (viviparity) as ichthyosaurs and mosasauroids (marine lizards) did^{4–6}. Viviparity has been proposed for plesiosaur, pachypleurosaur and nothosaur sauropterygians^{7–10}, but until now no concrete evidence has been advanced. Here we report two gravid specimens of *Keichousaurus hui* Young from the Middle Triassic of China. These exquisitely preserved specimens not only provide the first unequivocal evidence of reproductive mode and sexual dimorphism in sauropterygians, but also indicate that viviparity could have been expedited by the evolution of a movable pelvis in pachypleurosaurs. By extension, this has implications for the reproductive pattern of other sauropterygians and Mesozoic marine reptiles that possessed a movable pelvis.

Within the Sauropterygia, the two gravid specimens reported here are referred to the Pachypleurosauridae of the Pachypleurosauria on the basis of the presence of the following diagnostic features: the upper temporal opening smaller than the orbit, the presence of a distinct trough on the dorsal surface of the retro-articular process of the articular, no distal expansion of the sacral ribs, the pachyostotic pre- and postzygapophyses, and the reduced process-like dorsal iliac blade of the ilium². They can be further assigned to *Keichousaurus hui* Young, 1958 on the basis of the cervical region being longer than the trunk region, the short and blunt snout, the elongate upper temporal opening only slightly shorter than the orbit, the anterior position of the parietal opening, and the humerus longer and more robust than the femur⁹. Both specimens, housed in the National Museum of Natural Science (NMNS), Taichung, Taiwan, are nearly complete. One (NMNS-cyn2002-01) contains two embryos on each side, and the other (NMNS-VL191) at least three crushed embryos on each side. Both were collected from the Triassic limestone of Xingyi area, Guizhou province, southwestern China, probably from the late Middle Triassic Zhuganpo Member of the Falang Formation, as suggested by the fact that all other known specimens of this taxon have recently been confirmed to come from this member¹¹.

NMNS-cyn2002-01 is preserved in dorsal view. Its preserved length is about 296 mm, lacking the posterior portion of the tail after caudal 25 (Fig. 1a). The two embryos preserved on the right side are clearly more posterior in position than the two on the left, and at least the posterior one is directed posteriorly. The latter embryo, preserved in ventral view, reaches the second caudal rib posteriorly (cloacal or vent region) and the 11th dorsal (the 38th) vertebra anteriorly (Fig. 2). The anterior embryo, exposed in dorsal view, is not as well preserved, and is partly overlaid by the posterior embryo. The two left embryos are clearly separated from one another. The anterior one reaches anteriorly to the third dorsal

(the 30th) vertebra and the posterior one ends just anterior to the first sacral rib. The posterior one, exposed in ventral view, is clearly directed posteriorly. The anterior one is preserved in lateral view and seems to be directed anteriorly on the basis of the position of its scapula. The four embryos are mostly in articulation and their distribution on each side indicates that female *Keichousaurus hui* had a pair of oviducts as in ichthyosaurs¹² and many extant lizards¹³.

NMNS-VL191 has a total length of about 193 mm. It is preserved in dorsal view and compressed dorso-ventrally (Fig. 1b). It contains more embryos than NMNS-cyn2002-01 but they are poorly preserved. On the left side three embryos can be detected, all of which are posteriorly directed according to their rib orientation (Fig. 3). In contrast with NMNS-cyn2002-01, the left embryos of NMNS-VL191 are more posteriorly positioned than the right ones, posteriorly reaching the fourth caudal rib (although it may have been exaggerated by dorso-ventral compression *post mortem*), whereas the right embryos are more anteriorly placed, reaching the sixth dorsal (the 32nd) vertebra anteriorly. As in NMNS-cyn2002-01, the presence of embryos on both sides in NMNS-VL191 shows that female *Keichousaurus hui* must have had a pair of oviducts in life.

It has been well documented that embryos of ichthyosaurs are normally positioned head forwards as in modern cetaceans and that the gravid specimens with embryos head backwards represent an abnormal condition and might have caused the death of both mothers and embryos¹⁴. Most of the embryos in the two gravid specimens of *Keichousaurus hui* are head backwards and, similarly, they and their mothers might have been killed during birth because of abnormal carriage.

Although sexual dimorphism is common in the Pachypleurosauria^{8,9,15}, actual sex of each of the two morphotypes has never been demonstrated. In *Keichousaurus hui*, two sexes can be distinguished by the length ratio between the humerus and femur, and the structural complexity of the former. In one morph (sex X) the humerus is nearly as long as the femur and structurally simple, whereas in the other (sex Y) it is much longer than the femur and structurally massive⁹. In the two gravid specimens of *Keichousaurus hui*, the humerus lacks complicated structures and its length ratio to that of the femur is very similar to that of the specimens of sex X⁹. Therefore, in *Keichousaurus hui* sex X represents female and sex Y represents male. Because dimorphism in small European pachypleurosaurs such as *Neusticosaurus* and *Serpianosaurus*^{8,9,15} is closely

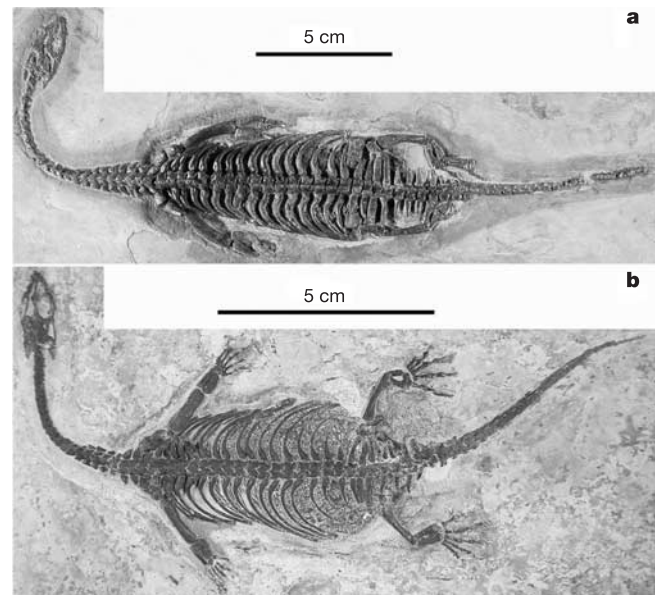


Figure 1 Two gravid specimens of *Keichousaurus hui* in dorsal view. **a**, NMNS-cyn2002-01; **b**, NMNS-VL191.

comparable to that seen in *Keichousaurus hui*, it is now possible to sex the two morphs in those taxa as well.

Primitive sauropterygians, such as pachypleurosaurians and nothosaurians, lived in near-shore environments, intraplatform basins and shallow epicontinental seas², and were capable of leaving the water to brood or bask¹⁶, indicating that either oviparity or viviparity could have been practical strategies for these marine reptiles. Therefore, the achievement of viviparity in these primitive sauropterygians could be confirmed only by direct fossil evidence such as the two current specimens. In contrast with those of extant egg-laying sea turtles and other marine reptiles, the sacral ribs of *Keichousaurus hui* are rod-like and never fused to the sacral vertebrae proximally, and their middle element form a unique, peg-and-socket-like joint with the reduced dorsal blade of the ilium

distally. This indicates that a chain-like connection is present between the pelvis and sacrum (Fig. 4a). As in other marine reptiles such as ichthyosaurs and mosasaurs, the loss of a solid connection between the pelvic girdle and sacrum is correlated with aquatic habits. It has been suggested that the absence of a firm sacro-iliac joint in *Keichousaurus hui* would allow relative movement, possibly to accommodate stress generated during a sudden stop or sharp turn as the animal swam⁹. Once evolved, this joint would also allow the pelvic girdle to change its shape, maximizing the space of the birth canal. A chain-like sacro-iliac joint would certainly enhance labour, allowing the live young to pass through the birth canal and emerge as quickly as possible in a vulnerable marine environment.

As in *Keichousaurus hui*, the sacral ribs did not fuse proximally with their respective vertebrae nor tightly articulate distally with

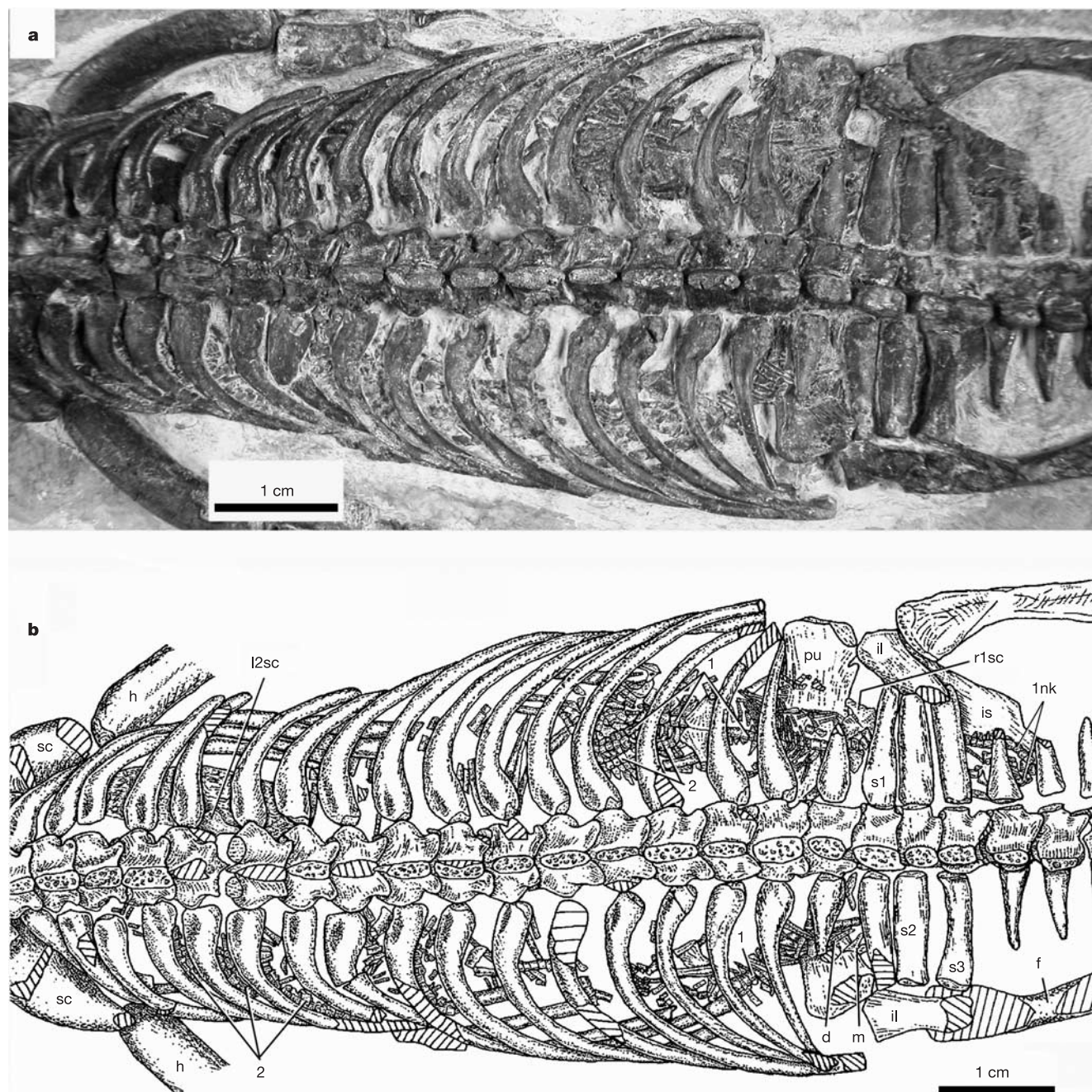


Figure 2 The trunk region of NMNS-cyn2002-01 in dorsal view. **a**, Actual specimen; **b**, drawing of **a**. Abbreviations: d, dentary; f, femur; h, humerus; is, ischium; l2sc, scapula of left anterior embryo; m, maxilla; pu, pubis of mother specimen; r1sc, scapula of right

posterior embryo; sc, scapula; s1–s3, sacral ribs 1–3; 1, posterior embryo of each side; 1nk, neck region of right posterior embryo; 2, anterior embryo of each side.

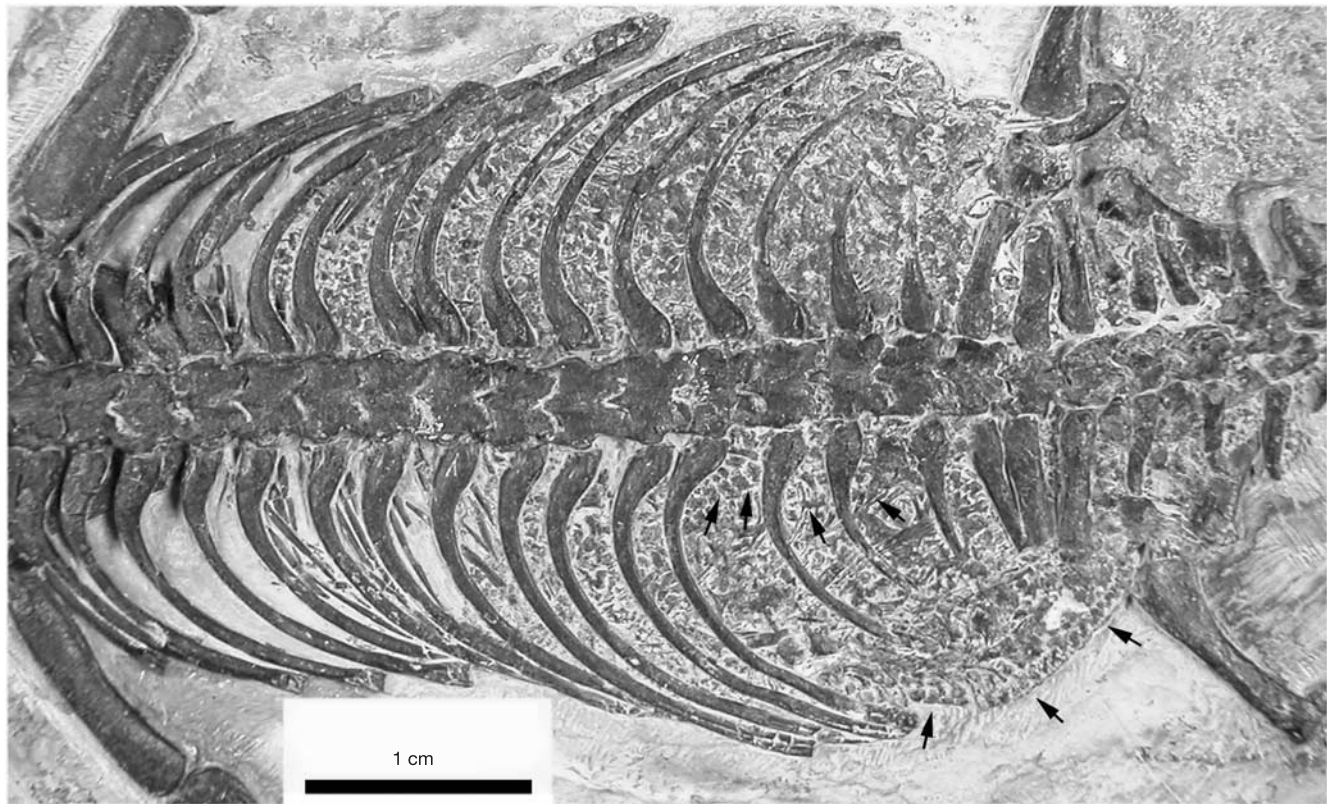


Figure 3 The trunk region of NMNS-VL191, showing three embryos (as indicated by arrows on left side).

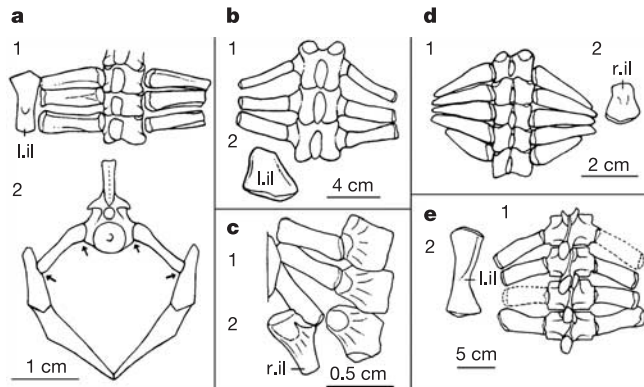


Figure 4 Sacral vertebrae and pelvis of selected sauropterygians.

a–c, Pachypleurosaur: **a**, *Keichousaurus hui* (1, the sacrum in dorsal view and left ilium in medial view, with its dorsal end directing downwards; 2, the reconstruction of 1 plus ventral part of the pelvic girdle, in posterior view; drawn based on NMNS-cyn2002-01); **b**, *Neusticosaurus* (= *Pachypleurosaurus*) *edwardsi* (1, the sacrum in dorsal view; 2, the left ilium in lateral view; redrawn from Fig. 19b in Carroll and Gaskill¹⁶); **c**, *Dactylosaurus gracilis* (= *schröderi*) (1, the sacrum in ventral and lateral view; 2, the right ilium in ventral view; redrawn from Fig. 1B in Sues and Carroll¹⁷). **d**, A nothosaur (*Lariosaurus valceresi*) (1, the sacrum in dorsal view; 2, the right ilium in medial view; drawn from Fig. 12B in Rieppel¹⁸). **e**, Plesiosaurs (1, the sacrum of *Cryptoclidus eurymerus* in dorsal view, redrawn from Fig. 14 in Brown²⁰; 2, the left ilium of *Bishanopliosaurus youngi* in medial view, redrawn from Fig. 6A in Sato *et al.*²⁵). Abbreviations: l.il, left ilium; r.il, right ilium; arrows indicate the movable (a chain-like) connection of the sacral ribs relative to the sacral vertebrae and to the pelvis, respectively.

the ilium in most other pachypleurosaur and nothosaur^{8,15–19}. This is generally interpreted as a phenomenon of pedomorphosis (juvenile condition)². In addition, the sacral ribs and the relevant elements of the pelvis such as the ilium were greatly simplified in

morphology, the former being bar-shaped and distally unexpanded or even tapering, whereas the dorsal blade of the latter became small and knob-like. These simplifications are comparable to those seen in *Keichousaurus hui*, suggesting that a similar mobility seen in the sacro-iliac joint of the former might have been present in those marine reptiles (Fig. 4b–d). If this is so, there was increased potential for the development of viviparity in these taxa as well. Furthermore, the pelvic girdle and sacrum also show a similar condition in advanced (derived) sauropterygians, such as plesiosaurs. To our knowledge, the sacral ribs are very simple in morphology and formed a weak connection with the bar-shaped ilium distally in adult plesiosaurs^{20–22} (Fig. 4e) although the fusion between the ribs and sacral vertebrae might occasionally occur in certain taxa (possibly in very old specimens)²³. Plesiosaurs, unlike pachypleurosaur and nothosaur, were considered to be the inhabitants of open seas²⁴. Their strong, well-developed forelimbs and hindlimbs are believed to be important in swimming, suggesting that a firm sacro-iliac joint would have been more appropriate. Yet the morphology of the sacral ribs and ilium indicates that the mobility might have been retained in articulations between the pelvis and sacrum. This indicates that plesiosaurs, like the aforementioned sauropterygians, might have been committed to viviparity. □

Received 30 June; accepted 23 September 2004; doi:10.1038/nature03050.

1. Carroll, R. L. *Vertebrate Paleontology and Evolution* (Freeman, New York, 1988).
2. Rieppel, O. *Encyclopedia of Paleoherpetology* Vol. 12A 1–134 (Verlag Dr Friedrich Pfeil, Munich, 2000).
3. Taylor, M. A. in *Ancient Marine Reptiles* (eds Callaway, J. M. & Nicholls, E. L.) xix–xlii (Academic, San Diego, 1997).
4. Johnson, R. Size independent criteria for estimating relative age and the relationships among growth parameters in a group of fossil reptiles (Reptilia, Ichthyosauria). *Can. J. Earth Sci.* **14**, 1916–1924 (1977).
5. Benton, M. J. The myth of the Mesozoic cannibals. *New Sci.* **12**, 40–44 (1991).
6. Caldwell, M. W. & Lee, M. S. Y. Live birth in Cretaceous marine lizards (mosasauroids). *Proc. R. Soc. Lond. B* **268**, 2397–2401 (2001).
7. Taylor, M. A. Lifestyle of plesiosaurs. *Nature* **319**, 179 (1986).
8. Sander, P. M. The pachypleurosaurids (Reptilia: Nothosauria) from the Middle Triassic of Monte San Giorgio (Switzerland), with the description of a new species. *Phil. Trans. R. Soc. Lond. B* **325**, 561–670 (1989).
9. Lin, K. & Rieppel, O. Functional morphology and ontogeny of *Keichousaurus hui* (Reptilia,

- Sauropterygia). *Fieldiana (Geol.)* **39**, 1–35 (1998).
10. Renesto, S., Lombardo, C., Tintori, A. & Danini, G. Nothosaurid embryos from the Middle Triassic of northern Italy: an insight into the viviparity of nothosaurs? *J. Vert. Paleontol.* **23**, 957–960 (2003).
 11. Wang, L. et al. Biostratigraphy of Triassic marine reptiles in southwest Guizhou and its adjacent area. *Acta Geol. Sinica* **74**, 349–353 (2001).
 12. Böttcher, R. Neue Erkenntnisse über die Fortpflanzungsbiologie der Ichthyosaurier (Reptilia). *Stuttgarter Beitr. Naturk. B* **164**, 1–51 (1990).
 13. Fox, H. in *Biology of the Reptilia* (ed. Gans, C.) 1–122 (Academic, London, 1977).
 14. Deeming, D. C., Halstead, L. B., Manabe, M. & Unwin, D. M. An ichthyosaur embryo from the Lower Lias (Jurassic: Hettangian) of Somerset, England, with comments on the reproductive biology of ichthyosaurs. *Mod. Geol.* **18**, 423–442 (1993).
 15. Rieppel, O. A new pachypleurosaur (Reptilia: Sauropterygia) from the Middle Triassic of Monte San Giorgio, Switzerland. *Phil. Trans. R. Soc. Lond. B* **323**, 1–73 (1989).
 16. Carroll, R. L. & Gaskill, P. The nothosaur *Pachypleurosaurus* and the origin of plesiosaurs. *Phil. Trans. R. Soc. Lond. B* **309**, 343–393 (1985).
 17. Sues, H.-D. & Carroll, R. L. The pachypleurosaurid *Dactylosaurus schroederi* (Diapsida: Sauropterygia). *Can. J. Earth Sci.* **22**, 1602–1608 (1985).
 18. Rieppel, O. The status of the sauropterygian reptile genera *Ceresiosaurus*, *Lariosaurus*, and *Silvestrosaurus* from the Middle Triassic of Europe. *Fieldiana* **38**, 1–46 (1998).
 19. Rieppel, O., Li, J. & Liu, J. *Lariosaurus xingyiensis* (Reptilia: Sauropterygia) from the Triassic of China. *Can. J. Earth Sci.* **40**, 621–634 (2003).
 20. Brown, D. S. The English Upper Jurassic Plesiosauroidea (Reptilia) and a review of the phylogeny and classification of the Plesiosauroidea. *Bull. Br. Mus. Nat. Hist. (Geol.)* **35**, 253–347 (1981).
 21. Bardet, N., Godefroit, P. & Sciau, J. A new elasmosaurid plesiosaur from the Lower Jurassic of southern France. *Palaeontology* **42**, 927–952 (1999).
 22. Sato, T. *Terminatorator ponteixensis*, a new elasmosaur (Reptilia: Sauropterygia) from the Upper Cretaceous of Saskatchewan. *J. Vert. Paleontol.* **23**, 89–103 (2003).
 23. Andrews, C. W. *A Descriptive Catalogue of the Marine Reptiles of the Oxford Clay, Part I* (Br. Mus. Nat. Hist., London, 1910).
 24. Sues, H.-D. Postcranial skeleton of *Pistosaurus* and interrelationships of the Sauropterygia (Diapsida). *Zool. J. Linn. Soc.* **90**, 109–131 (1987).
 25. Sato, T., Li, C. & Wu, X.-C. Restudy of *Bishanopliosaurus youngi* Dong 1980, a freshwater plesiosaurian from the Jurassic of Chongqing, China. *Vertebrata Palasiatica* **41**, 33–47 (2003).

Acknowledgements We thank R. Holmes and T. Sato of the Canadian Museum of Nature (CMN), Ottawa, for advice, suggestions and access to their reference collections, and R. Holmes for reading earlier drafts and for editorial assistance. X.-c. W. was supported by research grants from the CMN.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.-c.W. (xcwu@mus-nature.ca) or Y.-n. C. (joe@mail.nmns.edu.tw).

The evolution of alternative parasitic life histories in large blue butterflies

Thomas D. Als^{1,2,*}, Roger Vila^{2,*}, Nikolai P. Kandul², David R. Nash³, Shen-Horn Yen^{4,5}, Yu-Feng Hsu⁶, André A. Mignault², Jacobus J. Boomsma^{1,3} & Naomi E. Pierce²

¹Department of Ecology and Genetics, University of Aarhus, Ny Munkegade, 8000 Aarhus C, Denmark

²Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA

³Institute of Biology, Department of Population Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark

⁴Department of Biological Sciences, Imperial College at Silwood Park, Ascot, Berkshire, SL5 7PY, UK

⁵Department of Biological Sciences, National Sun Yat-Sen University, Kaohsiung 804, Taiwan

⁶Department of Life Sciences, National Taiwan Normal University, Taipei 116, Taiwan

*These authors contributed equally to this work

† Present address: Centre for Basic Psychiatric Research, Psychiatric Hospital in Aarhus, Skovagervej 2, 8240 Risskov, Denmark

Large blue (*Maculinea*) butterflies are highly endangered throughout the Palaearctic region, and have been the focus of intense conservation research^{1–3}. In addition, their extraordinary parasitic lifestyles make them ideal for studies of life history evolution. Early instars consume flower buds of specific host

plants, but later instars live in ant nests where they either devour the brood (predators), or are fed mouth-to-mouth by the adult ants (cuckoos). Here we present the phylogeny for the group, which shows that it is a monophyletic clade nested within *Phengaris*, a rare Oriental genus whose species have similar life histories^{4,5}. Cuckoo species are likely to have evolved from predatory ancestors. As early as five million years ago, two *Maculinea* clades diverged, leading to the different parasitic strategies seen in the genus today. Contrary to current belief, the two recognized cuckoo species show little genetic divergence and are probably a single ecologically differentiated species^{6–10}. On the other hand, some of the predatory morphospecies exhibit considerable genetic divergence and may contain cryptic species. These findings have important implications for conservation and reintroduction efforts.

Maculinea species have become the flagship butterflies for conservation in the UK and Europe^{2,11}. The severe decline of *Maculinea* populations during the twentieth century has been well documented, and all species have been included in the red data lists of most European countries³. Many management and reintroduction projects have been attempted, with variable results¹. The extinctions of the large blue (*Maculinea arion*) in the UK, the Netherlands and Belgium, the scarce large blue (*Maculinea teleius*) in the last two countries and the dusky large blue (*Maculinea nausithous*) in the Netherlands^{2,12}, have spurred increased conservation efforts, with large blue butterfly populations sometimes being used as bioindicators of habitat quality¹³.

Maculinea species are also the best-known examples of parasitic butterflies. Initially they feed on the flowers of specific Lamiaceae, Gentianaceae or Rosaceae host plants. When they reach the fourth instar, they drop to the ground and are picked up by *Myrmica*¹⁴ (or in a few cases *Aphaenogaster*¹⁵) ants and carried into the nest where they feed as parasites. Most currently recognized species, including the widely distributed species *M. arion*, *M. teleius* and *M. nausithous* and the east Asian *Maculinea arionides* prey on ant brood^{15–18}. In contrast, *Maculinea alcon* and *Maculinea rebeli* are ‘cuckoos’, whose larvae are fed primarily on regurgitations from ant workers, trophic eggs and prey items^{19,20}. These species have more elaborate adaptations of behavioural and chemical mimicry, and have thus been proposed to be derived relative to species that are strictly predatory²⁰. Several additional taxa occurring in the eastern Palaearctic have been proposed as species, including *Maculinea kurentzovi* and *Maculinea cyanecula*, but their status is still unclear and their life histories have not been described (see Supplementary Information)¹⁸.

More than 99% of the estimated 18,000 species of butterflies are herbivorous, but aphitophagy (carnivory and parasitism) has been fully documented in only about 80 species²¹. These are found primarily in the family Lycaenidae, to which *Maculinea* belongs, and are likely to be the result of the close relationship that the caterpillars of this family have with ants. Up to 75% of the approximately 5,000 species of Lycaenidae (*sensu stricto*) associate to some degree with ants, and whereas most of these relationships appear to be mutualistic, as many as 200 (4%) are known or suspected to be parasitic on ants²².

Maculinea belongs to the *Glaucompsyche* section of the Polyommataini, and *Sinia*, *Iolana*, *Caerulea* and *Phengaris* have been considered its closest relatives²³. In particular, the Oriental genus *Phengaris* has been proposed as the most likely sister group of *Maculinea* due to its similar morphology and the occurrence in the genus of both predatory and cuckoo parasitism on *Myrmica* ants^{4,5,24,25}. The present study reconstructs the evolution of ant parasitism, host plant association and speciation in *Maculinea*, and investigates whether the presently recognized species are likely to represent evolutionarily significant units for conservation.

Our molecular phylogeny includes 32 *Maculinea* specimens representing 31 geographically distinct populations of seven species

covering the entire Palaearctic from Denmark and Spain in the west to Japan and southeastern Russia in the east. We obtained 15 outgroup taxa, representing nine genera within the *Glaucopsyche* section²³. Maximum parsimony, maximum likelihood and bayesian methods (for each gene region and for all genes combined) were used to analyse information from 3,109 characters from mitochondrial *Cytochrome Oxidase I* and *II* (*COI* and *COII*) and nuclear *Elongation Factor1-alpha* (*EF1-α*) genes. Phylogenetic events were dated by applying published estimates of substitution rates for *COI* and *COII* to a phylogram recovered by maximum likelihood (see Methods).

Maculinea is recovered as a monophyletic group, with members of *Phengaris* as its closest relatives (Fig. 1). Although the criteria used to define a genus are largely subjective, the apparent paraphyly of *Phengaris* and the proximity of all three *Phengaris* specimens to *Maculinea* raises the possibility that the two genera should be synonymized under *Phengaris* Doherty, 1891, with the junior name *Maculinea* van Eecke, 1915 rendered invalid. The relationships among the remaining outgroup taxa are not well supported. Approximately five million years ago, *Maculinea* taxa separated into two main clades that correspond with the cuckoo and the predatory lifestyles (Fig. 2). The predatory clade is further divided into two

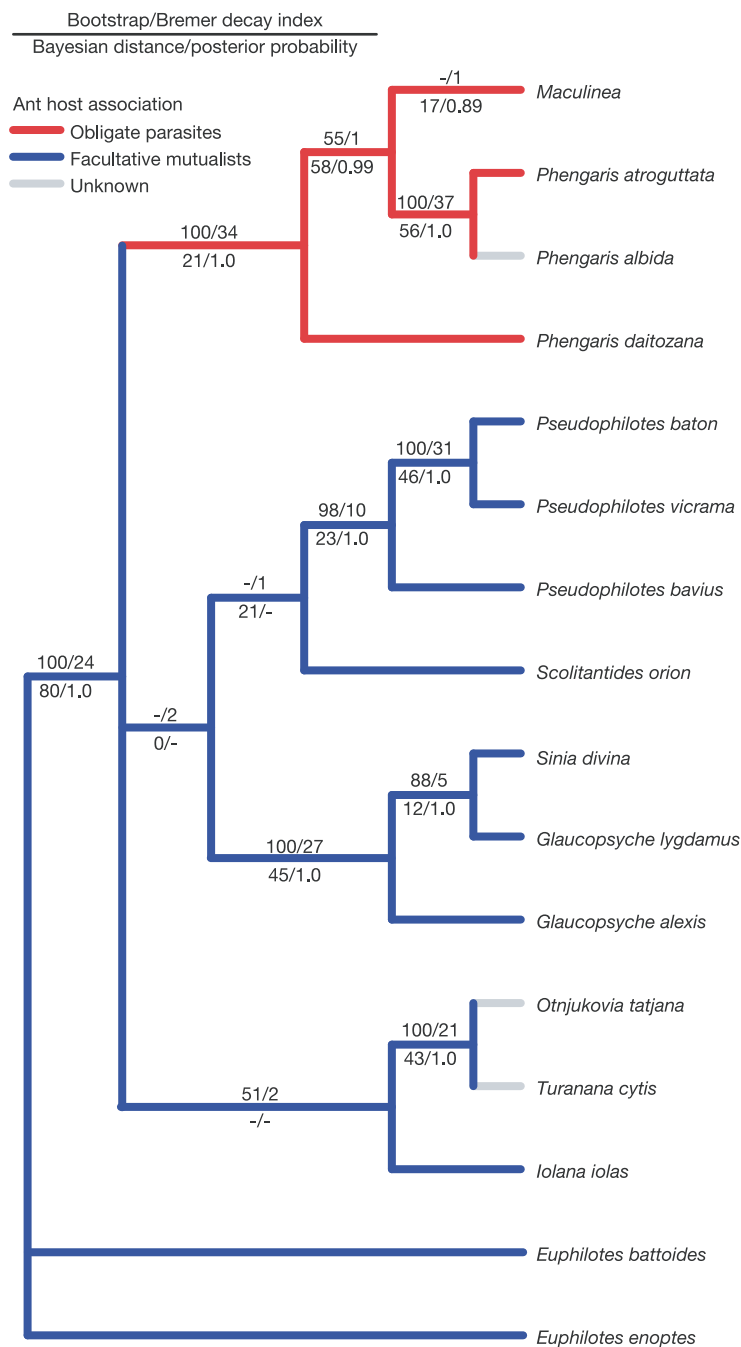
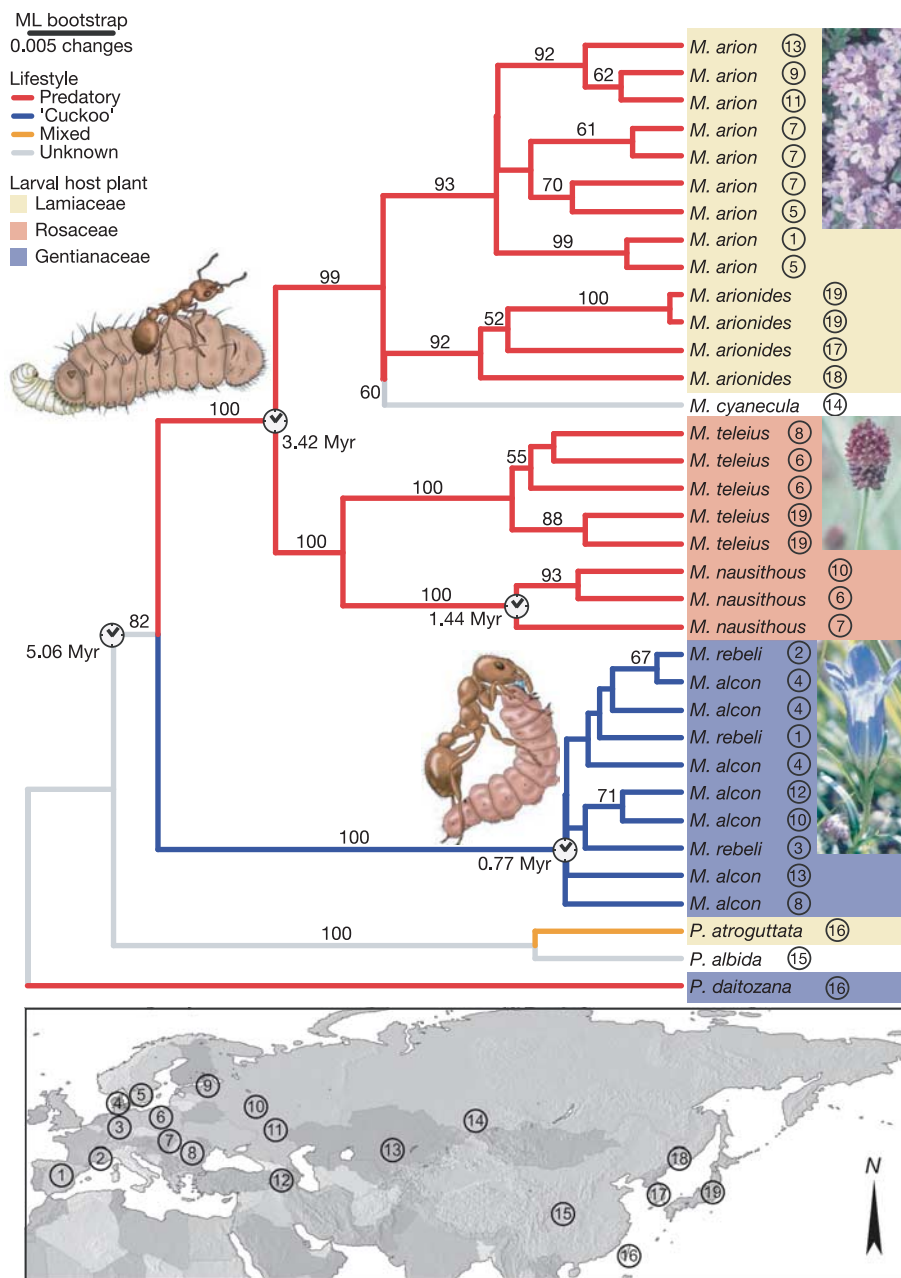


Figure 1 Phylogeny of the obligately parasitic genera *Phengaris* and *Maculinea* in relation to outgroup taxa that have facultative, mutualistic relationships with ants. The strict consensus maximum parsimony (MP) tree of 47 *Maculinea* and outgroup taxa was inferred from 3,109 base pairs (bp) of the genes *COI*, *COII* and *EF1-α*. The *Maculinea* part of the tree (32 specimens) is collapsed, and is shown in detail in Fig. 2. Photographs show

representative members of each genus except *Otrjukovia*. The strict consensus tree was constructed from 1,277 MP trees (tree length (TL) = 1,284; consistency index (CI) = 0.628; and retention index (RI) = 0.804). Bayesian inference recovered a similar topology.

Representatives of predatory species, especially *M. nausithous*, from different populations show considerable genetic divergence and may represent cryptic species (Supplementary Table 7). The cuckoo clade of *M. alcon* + *M. rebeli*, on the other hand, shows a distinct lack of structure, with a long, strongly supported basal branch and extremely limited differentiation among terminal branches (Fig. 2 and Supplementary Fig. 6). The minor extent of these differences, especially relative to the differences between populations of the predatory clades, suggests that *M. rebeli*'s status as a separate species is questionable. All extant local differentiation of *M. alcon* and *M. rebeli* between host plants and host ants^{6–8} may in fact derive from a single recent ancestor that arose considerably

Ancestral state reconstruction using maximum parsimony indicates that the ancestors of both the *Maculinea* clade and the *Phengaris* + *Maculinea* clade were butterflies from the eastern Palaearctic or Oriental region whose caterpillars mined the inflorescences of Gentianaceae or Lamiaceae during their initial instars and later parasitized *Myrmica* ants either as predators, or as mixed predators/cuckoos. Parsimony reconstruction does not distinguish between these alternative ancestral states, although both require fewer steps than a putative cuckoo ancestor without a predatory capacity. However, a maximum likelihood reconstruction taking



for *COL* in insects³⁰. This smoothed maximum likelihood phylogram of 35 taxa was inferred from 3,109 bp of the genes *COL*, *COLII* and *EF1-α* under the GTR + I + Γ nucleotide substitution model (−lnL = 4,655.43), with bootstrap values based on 500 pseudoreplicates.

into account branch lengths favours by 2 to 1 the predatory ancestral condition. This is consistent with most biological arguments for the evolution of feeding strategies in this group^{19,20} (Supplementary Figs 8 and 9, Supplementary Table 8).

The life history of the most basal ant-parasitic species, *Phengaris daitozana*^{4,5,24}, might represent that of a hypothetical predatory ancestor. Closely related *Glaucopteryx* larvae sometimes over-winter as pupae in ant nests, presumably because of the stability and security afforded by the nest^{26,27}. The life history shown by *P. daitozana* may have evolved simply through a shift from pupal to larval diapause^{4,5,24}, necessitating additional feeding in the spring before pupation. *P. daitozana* completes its nutrition at this point by feeding on ant brood. Larvae of the more derived *Phengaris atroguttata*, however, after a similar initial bud-mining phase on the flowers of Lamiaceae, are actively carried into the nest by *Myrmica* workers, where they feed both as cuckoos and as predators of the ant brood^{4,5,24}. Because our study includes all species of *Maculinea* and *Phengaris* with known life histories, and our conclusions are not affected by the paraphyly or monophyly of *Phengaris*, new biological observations will be necessary to reach a better understanding of the evolution of parasitic strategies in this group.

Phylogenetic conservatism in host ant use is exhibited at the subfamily level: *Maculinea* + *Phengaris* form a clade whose species parasitize the ant subfamily Myrmicinae, almost exclusively the genus *Myrmica*. Similarly, all of the described species in the related polyommata genus *Lepidochrysops*, whose independently evolved phyto-parasitic lifestyle appears to have been associated with significant diversification (about 120 species), parasitize ants in the subfamily Formicinae, primarily the genus *Camponotus*^{22,26}. Other groups of parasitic Lycaenidae show similar degrees of specialization, with four species of Liphyrini all attacking weaver ants, *Oecophylla* (Formicinae), and an estimated 27 species of *Thestor* and five species of *Trimenia* associating with pugnacious ants, *Anoplolepis* (Formicinae)²². A recent phylogeny of the Australian butterfly genus *Acrodipsas* shows a shift in which the most basal species feed on ants in the subfamily Dolichoderinae, whereas all known derived species parasitize ants in the Myrmicinae²⁸.

Detailed population-level information about host ant use is only available for *Maculinea* species in Europe^{7,10,14,20}. Here, the host ant species used by *Maculinea* have been shown to be of prime importance for conservation^{1,2,11} and reintroduction programmes^{1,2}. Between them, the five European *Maculinea* species use host ants from all of the species-groups of free-living *Myrmica* found on the continent²⁹. However, any single population of *Maculinea* normally depends on only one or two host ant species^{6–8,14,20}, and can show considerable local adaptation to its hosts^{6–8}. Although they comprise the most genetically homogeneous clade in the phylogeny, *M. alcon* + *M. rebeli* use a total of seven *Myrmica* species as major hosts within Europe (Supplementary Table 10). The relatively broad range in host ant species together with the observation that individual populations are typically highly specific with respect to ant association suggest that cuckoo taxa may be undergoing rapid ecological divergence⁶. In contrast, the most genetically divergent species, the predatory *M. nausithous*, uses only two host ant species across its entire recorded range (Supplementary Table 10).

The present results overturn widely held inferences about *Maculinea* evolution and conservation. Future conservation and reintroduction programmes of predatory *Maculinea* species will need to take the possibility of cryptic species into account, and would thus benefit from genetic screening of alternative source populations. In contrast, the recognition and conservation of evolutionarily significant units for cuckoo species will need to be based on persistent ecological and behavioural adaptations of local populations. Finally, the now extremely sparse and threatened populations of Oriental *Phengaris* species should receive high conservation priority to enable further study of the evolutionary origins of the unusual parasitic life histories in this clade. □

Methods

Methods are described in a more detailed manner, and are fully referenced, in the Supplementary Information.

Specimen sequences

Total genomic DNA was extracted from 47 specimens of *Maculinea* and outgroup taxa. Two mitochondrial genes, *COI* and *COII*, and a nuclear gene, *EF1-α* were amplified by polymerase chain reaction (PCR). The direct sequencing of double-stranded PCR products yielded fragments of equal lengths, and alignments were unambiguous for both mitochondrial and nuclear gene fragments.

Phylogenetic analysis

Phylogenetic analyses were conducted using maximum parsimony (MP; PAUP* 4.0b10), maximum likelihood (ML; PHYML) and bayesian inference (BI; MrBayes 2.01). Monophyly of *Maculinea* and *Phengaris* + *Maculinea* was inferred using the total data set of 47 specimens, whereas relationships within *Maculinea* were inferred using a subset consisting of 32 specimens of *Maculinea* and three of *Phengaris* as outgroups to minimize potential for long-branch attraction. Both separate and combined phylogenetic analyses of mitochondrial and nuclear genes were performed. Hierarchical likelihood ratio tests (hLRTs) were used to determine the best-fitted substitution model for each data set in ML and BI analyses. To ensure that BI was not trapped in local optima, each BI analysis was run three times, and average log-likelihood (lnL) values at stationarity were calculated and compared for convergence. Nonparametric bootstrap values were used to estimate the support of tree branches recovered by MP and ML.

Dating main phylogenetic events

An ML phylogram recovered from the *COI* + *COII* data set of 32 *Maculinea* and three *Phengaris* specimens in PHYML was used to date main phylogenetic events. The likelihood ratio test (LRT) found a significant deviation from substitution rate consistency ($P < 0.0001$) across different branches on the ML topology. A nonparametric rate-smoothing (NPRS) algorithm was therefore used to homogenize evolutionary rates across the topology. The topology was then calibrated by applying a published estimate of substitution rate³⁰ in *COI* to the mean uncorrected pairwise distance for the single calibration node.

Ancestral character state reconstruction

Ancestral character state analyses were performed using Mesquite v. 1.01. Both MP and ML character optimizations were applied to the ML phylogram for 47 taxa inferred from the combined analysis of *COI* + *COII* and *EF1-α* genes under the GTR + I + Γ model of DNA substitution (Supplementary Fig. 6) trimmed to include only one sample for each of the *Phengaris* and *Maculinea* species. For MP analyses, host plant families were coded as a multistate unordered character. Two hypotheses for the parasitic strategy were considered: coding the *Phengaris* and *Maculinea* cuckoo strategies as equivalent, or as one-step different. Two coding options were used to cover all possibilities and explore the implications of both models. ML optimizations were done using the Markov k-state one-parameter model.

Received 16 May; accepted 14 September 2004; doi:10.1038/nature03020.

- Munguia, M. L. & Martin, J. *Action Plan for Maculinea Butterflies in Europe* (Council of Europe Publishing, Strasbourg, 1999).
- Thomas, J. A. in *Ecology and Conservation of Butterflies* (ed. Pullin, A. S.) 180–197 (Chapman & Hall, London, 1995).
- Van Swaay, C. A. M. & Warren, M. S. *Red Data Book of European Butterflies (Rhopalocera)* 124–139 (Council of Europe Publishing, Strasbourg, 1999).
- Uchida, H. *The Beautiful Butterfly Island Formosa Forever* (Shinwa, Tokyo, 1995).
- Jean, J. L. *Ecological Studies on the Two Myrmecophilous Butterflies Phengaris* (Department of Plant Pathology and Entomology, National Taiwan University, Taipei, 1996).
- Als, T. D., Nash, D. R. & Boomsma, J. J. Adoption of parasitic *Maculinea alcon* caterpillars (Lepidoptera: Lycaenidae) by three *Myrmica* ant species. *Anim. Behav.* **62**, 99–106 (2001).
- Als, T. D., Nash, D. R. & Boomsma, J. J. Geographical variation in host-ant specificity of the parasitic butterfly *Maculinea alcon* in Denmark. *Ecol. Entomol.* **27**, 403–414 (2002).
- Elmes, G. W., Akino, T., Thomas, J. A., Clarke, R. T. & Knapp, J. J. Interspecific differences in cuticular hydrocarbon profiles of *Myrmica* ants are sufficiently consistent to explain host specificity by *Maculinea* (large blue) butterflies. *Oecologia* **130**, 525–535 (2002).
- Gadeberg, R. M. E. & Boomsma, J. J. Genetic population structure of the large blue butterfly *Maculinea alcon* in Denmark. *J. Insect Conserv.* **1**, 99–111 (1997).
- Steiner, E. M. et al. Host specificity revisited: New data on *Myrmica* host ants of the lycaenid butterfly *Maculinea rebeli*. *J. Insect Conserv.* **7**, 1–6 (2003).
- Elmes, G. W. & Thomas, J. A. Complexity of species conservation in managed habitats: interaction between *Maculinea* butterflies and their ant hosts. *Biodivers. Conserv.* **1**, 155–169 (1992).
- Wynhoff, I. The recent distribution of the European *Maculinea* species. *J. Insect Conserv.* **2**, 15–27 (1998).
- Geisler-Strobel, S. Landscape planning orientated studies on ecology, distribution, endangerment and protection of the lycaenids *Glaucopteryx* (*Maculinea*) *nausithous* and *Glaucopteryx* (*Maculinea*) *teleius*. *N. Entomol. Nach.* **44**, 1–105 (1999).
- Thomas, J. A., Elmes, G. W., Wardlaw, J. C. & Woyciechowski, M. Host specificity among *Maculinea* butterflies in *Myrmica* ant nests. *Oecologia* **79**, 452–457 (1989).
- Yamaguchi, S. *The Life Histories of Five Myrmecophilous Lycaenid Butterflies of Japan* 21–116 (Kodansha, Tokyo, 1988).
- Elferich, N. W. New facts on the life history of the dusky large blue *Maculinea nausithous* (Lepidoptera: Lycaenidae) obtained by breeding with *Myrmica* ants in plaster nests. *DEINSEA* **4**, 97–102 (1998).
- Thomas, J. A. & Wardlaw, J. C. The capacity of a *Myrmica* ant nest to support a predacious species of *Maculinea* butterfly. *Oecologia* **91**, 101–109 (1992).

18. Sibatani, A., Saigusa, T. & Hirowatari, T. The genus *Maculinea* van Eecke, 1915 (Lepidoptera: Lycaenidae) from the East Palaearctic Region. *Tyô to Ga* **44**, 157–220 (1994).
19. Elmes, G. W., Thomas, J. A. & Wardlaw, J. C. Larvae of *Maculinea rebeli*, a large-blue butterfly, and their *Myrmica* host ants: wild adoption and behaviour in ant-nests. *J. Zool.* **223**, 447–460 (1991).
20. Thomas, J. A. & Elmes, G. W. Higher productivity at the cost of increased host specificity when *Maculinea* butterfly larvae exploit ant colonies through trophallaxis rather than by predation. *Ecol. Entomol.* **23**, 101–108 (1998).
21. Pierce, N. E. Predatory and parasitic Lepidoptera: Carnivores living on plants. *J. Lep. Soc.* **49**, 412–453 (1995).
22. Pierce, N. E. *et al.* The ecology and evolution of ant association in the Lycaenidae (Lepidoptera). *Annu. Rev. Entomol.* **47**, 733–771 (2002).
23. Eliot, J. N. The higher classification of Lycaenidae (Lepidoptera): a tentative arrangement. *Bull. Br. Mus. Nat. Hist. (Entomol.)* **28**, 371–505 (1973).
24. Igarashi, S. & Fukuda, H. *The Life Histories of Asian Butterflies* Vol. II 552–555 (Tokai Univ. Press, Tokyo, 2000).
25. Fiedler, K. Lycaenid–ant interactions of the *Maculinea* type: tracing their historical roots in a comparative framework. *J. Insect Conserv.* **2**, 3–14 (1998).
26. Fiedler, K. Systematic, evolutionary, and ecological implications of myrmecophily within the Lycaenidae (Insecta: Lepidoptera: Papilionoidea). *Bonn. Zool. Monogr.* **31**, 1–210 (1991).
27. Tilden, J. W. An occurrence of pupa of *Glaucopsyche lygdamus behrii* (Edwards) in an ant nest (Lepidoptera: Lycaenidae). *Pan-Pac. Entomol.* **23**, 42–43 (1947).
28. Eastwood, R. & Hughes, J. M. Molecular phylogeny and evolutionary biology of *Acrodipsas* (Lepidoptera: Lycaenidae). *Mol. Phylogenet. Evol.* **27**, 93–102 (2003).
29. Radchenko, A. G. Taksonomicheskaya struktura roda *Myrmica* (Hymenoptera, Formicidae) Evrazii. [Taxonomic structure of the genus *Myrmica* (Hymenoptera, Formicidae) in Eurasia.]. *Zool. Zh.* **73**, 39–51 (1994).
30. Quek, S. P., Davies, S. J., Itino, T. & Pierce, N. Codiversification in an ant–plant mutualism: the phylogeny of host use in *Crematogaster* (Formicidae) associates of *Macaranga* (Euphorbiaceae). *Evolution* **58**, 554–570 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Aoki, S. C. Collins, J. T. Costa, J. Dantart, A. V. Dantchenko, H. Deng, V. Dinca, K. Dunn, G. W. Elmes, K. Emura, E. Figurny-Puchalska, M. Goia, D. Hsu, H.-C. Huang, T. Inomata, Y.-J. Jang, J. L. Jean, J. Jubany, U. Kurosu, D. J. Lohman, V. A. Lukhtanov, T. Mizusawa, P. S. Nielsen, M. Owada, R. Rougerie, A. Shapiro, C. Stefanescu, H. Sugiyama, M. W. Tan, J. A. Thomas, M. A. Travassos, S. Viader, J. C. Wardlaw, S. Yamaguchi and D. Zitzman for providing material. D. K. Aanen, R. Eastwood, B. D. Farrell, G. E. Morse, D. B. Rand, J. A. Thomas, A. S. Sequeira and P. Villesen gave suggestions and comments, and P. Ackery allowed us to use images from the BMNH. Support for this research was provided by a National Science Foundation grant to N.E.P., a Danish Natural Science Research council grant to J.J.B., a Fulbright Postdoctoral Fellowship to R.V., grants from the Putnam Expeditionary Fund of the Museum of Comparative Zoology to N.E.P., N.P.K. and R.V., the Natural Conservation Foundation, Council of Agriculture and National Science Council, Taiwan, R.O.C. to Y.F.H., a PhD student fellowship from Harvard's Department of Organismic and Evolutionary Biology to N.P.K., a PhD student fellowship from the University of Aarhus to T.D.A., and the EU RTD network 'MacMan'.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.E.P. (npierce@fas.harvard.edu). The sequences have been deposited in GenBank under accession numbers AY675355–AY675448.

An obligate brood parasite trapped in the intraspecific arms race of its hosts

Bruce E. Lyon¹ & John McA. Eadie²

¹Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064, USA

²Department of Wildlife, Fish and Conservation Biology, University of California, Davis, California 95616-8751, USA

Reciprocal selection pressures often lead to close and adaptive matching of traits in coevolved species. A failure of one species to match the evolutionary trajectories of another is often attributed to evolutionary lags^{1,2} or to differing selection pressures across a geographic mosaic^{3,4}. Here we show that mismatches in adaptation of interacting species—an obligate brood parasitic duck and each of its two main hosts—are best explained by the

evolutionary dynamics within the host species. Rejection of the brood parasite's eggs was common by both hosts, despite a lack of detectable cost of parasitism to the hosts. Egg rejection markedly reduced parasite fitness, but egg mimicry experiments revealed no phenotypic natural selection for more mimetic parasitic eggs. These paradoxical results were resolved by the discovery of intraspecific brood parasitism and conspecific egg rejection within the hosts themselves. The apparent arms race between species seems instead to be an incidental by-product of within-species conflict, with little recourse for evolutionary response by the parasite.

Avian obligate brood parasites depend entirely on other species to raise their offspring, often inflicting severe fitness costs on hosts. Brood parasitism provides a model system for investigating the dynamics of antagonistic coevolution, because of the reciprocally hostile relationship between parasite and host^{1,2,5–8}. In some parasitic taxa, extreme fitness costs of parasitism to hosts have favoured the evolution of egg discrimination and rejection by hosts, which in turn has led to the evolution of egg mimicry and host specialization in the parasite^{5–8}. The black-headed duck (*Heteronetta atricapilla*) of southern South America is unique in comparison with all other species of obligate brood parasites in that its highly precocial chicks leave the host nest within a day of hatching (Fig. 1g) and require no post-hatching parental care^{9,10}. This parasite should impose few fitness costs on its hosts and, accordingly, the ecological and evolutionary dynamics of host–parasite interactions should differ markedly from those of all other brood parasites.

We conducted a large-scale observational and experimental study of host–parasite interactions in black-headed ducks during four breeding seasons on seven wetlands in the pampas of Argentina. Brood parasitism was common (29.3% of 1,927 potential host nests of 11 species parasitized). Several attributes of the brood parasitism were counter to those expected for a precocial brood parasite. First, the parasites used very few host species (Fig. 1a), and parasitized these hosts at a high frequency (Fig. 1b). Despite the diversity of species used at least occasionally as hosts in our study (11 species), 80% of the 974 duck eggs we found occurred in nests of just two species of coots (Fig. 1a, e, f), with almost half occurring in a single host, the red-gartered coot (*Fulica armillata*). Because hatching success of the duck eggs is highest with this host (Fig. 1c), an estimated 58% of all ducklings hatch from nests of this one species and 83% from both coot species combined (see Methods). Dependence on such a narrow range of hosts was unexpected because the ability to use a wide diversity of hosts has been proposed as a key factor in the evolution of obligate brood parasitism in *Heteronetta*^{2,10}. Second, the parasitic eggs had low hatching success in both main hosts (Fig. 1c), despite similar incubation periods of host and parasite. Third, both main hosts showed high levels of egg rejection (Fig. 1d). A strong negative correlation between the frequency of egg rejection and the hatching success of duck eggs for each host on each wetland (Spearman rank correlation $r_s = -0.99$, $n = 8$, $P < 0.01$) indicates that egg rejection markedly decreases the reproductive success of black-headed ducks and is a main source of egg mortality.

Egg rejection has arisen independently in a wide variety of birds to counter the costs of interspecific brood parasitism^{1,2,5–8}; its occurrence here therefore implies some cost of parasitism to hosts. Such costs would have to be borne during incubation because the ducklings leave the nest within a day of hatching. Using both naturally and experimentally parasitized nests, we assessed costs known to be suffered by hosts of brood parasitism, including smaller host clutch size, longer incubation period, increased egg loss¹¹ and increased nest predation risk from the non-cryptic duck eggs¹² (Fig. 1h). We detected no costs of parasitism for red-gartered coots, whereas parasitized red-fronted coot (*F. rufifrons*) nests suffered higher egg loss rates than unparasitized nests (Table 1). Whereas many costs of parasitism are reduced by, and thus select for, egg rejection, 'unrecoverable' costs—such as incidental egg

displacement or damage by the brood parasite during parasitism—do not¹³. Further analysis of egg loss in parasitized red-fronted coot nests based on experimental egg addition and removal nests (see Methods) revealed that egg loss is associated only with the act of parasitism itself, not the presence of parasitic eggs, a cost that would not promote the evolution of egg rejection (Table 1).

We did not directly measure the energetic costs to incubating parents of caring for parasitic eggs, which could be an important fitness cost to hosts¹⁴. However, in the closely related American coot (*F. americana*), a species with larger clutch sizes, fat reserves actually increase throughout incubation, indicating that incubation might not be energetically costly in coots¹⁵.

Independently of the factors selecting for egg rejection by hosts, the strong impact of egg rejection on the fitness of the black-headed duck should select for counter-adaptations such as the evolution of egg mimicry, particularly because the parasites depend on very few host species. We used field experiments with painted hen or host eggs (see Methods) to determine whether incremental improvement in egg mimicry (mimicking background colour alone or both the colour and shape of host eggs; Fig. 2a) would enhance the

acceptance rate, and hence the hatching success, of duck eggs in host nests. The degree of mimicry did not affect egg rejection rates in either host species (Fig. 2b, c); all egg treatments were rejected at similar rates and at rates within the range observed for real black-headed duck eggs.

Our results present two findings that are inconsistent with the hypothesis that interspecific interactions have driven host–parasite evolution in this system: high levels of egg rejection by the hosts in the absence of detectable costs of parasitism, and egg rejection that does not favour natural selection for egg mimicry in the brood parasite, at least over the range of egg features we examined. However, our results are consistent with an alternative hypothesis; specifically, that egg rejection evolved as a mechanism to reduce the costs of intraspecific brood parasitism within the host populations, and that rejection of duck eggs is an incidental by-product of this mechanism. A similar explanation has been proposed for egg rejection in weaverbirds (Ploceidae)^{16,17}, but the influence of interspecific parasitism has not yet been assessed^{2,18}.

We found that intraspecific brood parasitism occurs regularly in both species of coots: females laid eggs in the nests of conspecifics, and hosts recognized and rejected some of these conspecific parasitic eggs. We studied intraspecific parasitism in red-gartered coots in 1997 and determined that at least 13% of 266 nests were parasitized by conspecifics. Retrospective analysis of our census data from previous years revealed average detectable rates of intraspecific parasitism of 4.7% in 254 red-gartered coot nests (range for individual wetlands 2.7–12.9% of nests) and 5.2% in 212 red-fronted coot nests (range for individual wetlands 2.9–7.9% of nests). These rates are considerable underestimates given that our earlier studies were not focused on detecting intraspecific parasitism (see Methods). Nine of 35 (26%) red-gartered coots rejected at least one parasitic coot egg, and 6 of 23 birds (26%) rejected conspecific eggs that we added experimentally to their nests, indicating that hosts are capable of sophisticated egg discrimination that goes well beyond distinguishing between duck and coot eggs. Red-fronted coots are also capable of recognizing and rejecting conspecific eggs: parasitic eggs were rejected at two of the nine (12%) parasitized nests.

Conspecific egg rejection is rare in birds¹⁹ and seems difficult to evolve; its presence in the two hosts of black-headed ducks is difficult to explain other than as a defence against the costs of conspecific brood parasitism. Intraspecific brood parasitism and egg rejection are widespread in the rail family (Rallidae), including several other species of coots^{19–21}, none of which are parasitized by interspecific brood parasites. Detailed studies of the American coot

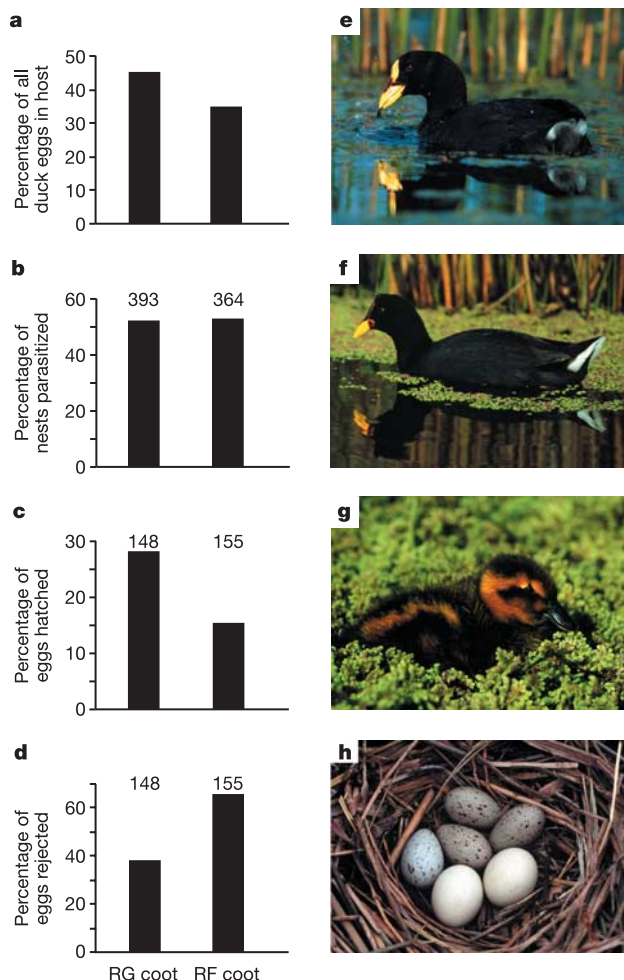


Figure 1 Frequency and attributes of parasitism in the two main hosts of the black-headed duck, the red-gartered coot (RG coot) and red-fronted coot (RF coot).

a, Percentage of the total 974 duck eggs encountered during the study that were laid in nests of the two main hosts. **b**, Percentage of nests of each species parasitized by the ducks. **c**, Percentage of duck eggs laid in nests of each host species that hatched. **d**, Percentage of the duck eggs laid in nests of each host species that were rejected. The sample size above each bar indicates the number of nests (**b**) or eggs (**c**, **d**). **e**, Red-gartered coot. **f**, Red-fronted coot. **g**, One-day-old black-headed duckling, the age at which the ducklings become completely independent of hosts. **h**, Parasitized red-gartered coot nest with two duck eggs, showing a lack of mimicry.

Table 1 Potential costs of parasitism by black-headed ducks on two main host species

Parameter	Parasitized	Not parasitized	P
Red-gartered coot			
Clutch size	4.10 ± 0.09 (144)	4.11 ± 0.09 (151)	0.95*
Host incubation period (d)	22.5 ± 1.24 (11)	23.4 ± 0.65 (11)	0.50*
Nests losing host eggs (%)	16.4 (61)	11.5 (87)	0.39†
Nest predation rate (%)	27.9 (86)	26.5 (49)	0.86†
Red-fronted coot			
Clutch size	5.04 ± 0.14 (53)	5.20 ± 0.14 (56)	0.44†
Host incubation period (d)	21.5 ± 0.89 (9)	23.1 ± 0.60 (22)	0.15*
Nests losing host eggs§ (%)	28.7 (94)	9.8 (61)	< 0.005†
Nest predation rate (%)	25.0 (44)	35.4 (48)	0.29†

Statistical tests for comparison of parasitized and unparasitized nests: *Student's *t*-test; † χ^2 test; ‡analysis of covariance to control for seasonal influence on clutch size; means are least-square means.

§We used a combination of experimental and observational nests to distinguish egg loss due to displacement or damage during the act of parasitism (a cost not affected by egg rejection) from egg loss due to the presence of duck eggs after parasitism (a cost prevented by early egg rejection). We experimentally removed duck eggs from parasitized nests to assay the former, and we experimentally added duck eggs to unparasitized nests to assay the latter. We then combined these experimental nests with the observational nests and used logistic regression to partition the two sources of egg loss statistically: the act of parasitism itself affected host egg loss (logistic regression; Wald χ^2 for parasitized versus unparasitized nests = 7.86, $P = 0.005$), but the presence of duck eggs did not (Wald χ^2 for the duration for which parasitic eggs are in a nest = 0.003, $P = 0.96$). Data are shown as means ± s.e.m. or as percentages, with the sample size in parentheses.

in North America reveal high costs of intraspecific brood parasitism to hosts and confirm that egg rejection is an evolutionary response specifically to reduce these costs¹⁹. Similar costs are likely to apply to the two Argentine coots, given their high chick provisioning rates (B.E.L. and J.McA.E., personal observation) and the fact that both species have highly ornamented chicks, a characteristic of extreme competition for limited food in American coots²².

The switch of perspective from between-species to within-host dynamics can explain why the hosts reject duck eggs, even though parasitism by black-headed ducks seems not to be costly to them. It can also explain why more mimetic duck eggs are not favoured: given that hosts have been selected to distinguish their own eggs from those of other coots, the duck eggs differ too much from host eggs for incremental changes in appearance to increase their acceptance rate (Fig. 1h). Perfectly mimetic eggs would be selectively favoured over the existing white eggs, because both host species rejected conspecific parasitic eggs at a lower rate than the experimental mimicry eggs (all treatments combined; Fisher's exact $P < 0.01$ for both species) or real duck eggs ($P < 0.05$ for red-gartered coots; $P < 0.01$ for red-fronted coots). However, because incremental changes in shape and background colour do not improve the acceptance rate of duck eggs (Fig. 2), several simultaneous, independent changes in egg features would be needed to achieve such sophisticated mimicry. Furthermore, given that all members of the waterfowl order Anseriformes have immaculate eggs²³, the evolution of spotted eggs is likely to be phylogenetically

constrained. Black-headed ducks seem to be trapped in the social conflict of their hosts, without recourse to evolutionary counter-adaptations, at least with regard to the possibility of reducing egg rejection rates.

If hosts can recognize subtle differences between conspecific eggs, should not the strikingly different duck eggs always be rejected? An experiment conducted 30 years ago for other reasons demonstrates clearly that acceptance of non-mimetic foreign eggs can occur even where selection has acted only on host recognition and rejection of conspecific eggs. Weller added experimental white hen eggs to American coot nests in Iowa²⁴, a species with high frequencies of intraspecific brood parasitism and egg rejection¹⁹. Because this species does not suffer any interspecific brood parasitism, all aspects of egg recognition and rejection must stem from selection by intraspecific brood parasitism. As with their South American relatives, American coots rejected some, but not all, of the hen eggs added to their nests; 44% of 27 eggs added to nests on stable wetlands were accepted and incubated by the host. The striking similarity between the results of Weller's experiments and ours supports the hypothesis that rejection of duck eggs by the two South American coots is an incidental by-product of social strife within the hosts themselves.

The cognitive and ecological factors influencing the partial acceptance of duck eggs by coots remain unclear. One possibility is that young coots breeding for the first time learn to recognize their own eggs through imprinting, so that individuals parasitized during their first nesting attempt imprint on both their own and duck eggs, becoming acceptors for their entire lifetime^{2,25,26}. This hypothesis is rejected by our observation that all individuals in both host species seem capable of recognizing duck eggs, even though they do not always reject them. During floods (both species) or wind-driven high waves in one open wetland (red-gartered coot), rejection rates increased to 100%. Weller found an identical pattern with American coots—in nests subjected to severe flooding, coots rejected 100% of experimental hen eggs²⁴.

An alternative possibility is that the costs of rejection (that is, rejection of the host's own eggs²⁷) are state-dependent, such that different individuals show different degrees of rejection behaviour. Detailed study of the mechanisms of both egg recognition and rejection is now required for an understanding of how these enigmatic brood parasites are able to obtain a sufficient level of egg acceptance to persist. Indeed, rather than the generalist brood parasite once envisaged^{2,10}, black-headed ducks might instead be exploiting a rather narrow niche defined by the cognitive limits of their two main hosts. □

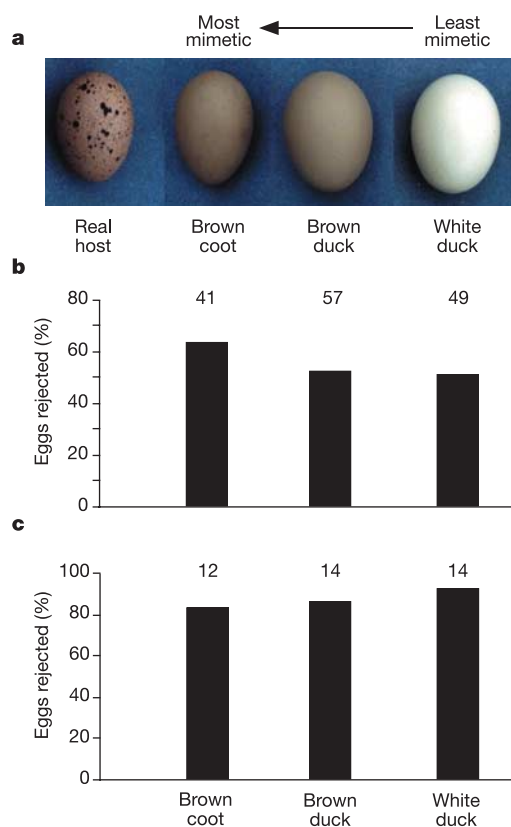


Figure 2 Results of egg mimicry experiments in the two principal hosts. **a**, Examples of a real host egg and the mimetic series of model eggs used in the experiment, arranged from most mimetic (brown coot) to least mimetic (white duck). The white duck treatment was similar in appearance and shape to real duck eggs. **b**, **c**, The degree of mimicry did not affect egg rejection rates by red-gartered coots (**b**) or by red-fronted coots (**c**). With degree of mimicry entered as a ranked variable, logistic regressions revealed no effect of mimicry rank on the proportion of eggs rejected for either red-gartered coots (Wald $\chi^2 = 0.98$, $P = 0.32$) or red-fronted coots (Wald $\chi^2 = 0.55$, $P = 0.46$). The sample size above each bar indicates the number of nests in each treatment.

Methods

Detecting and monitoring parasitism

The biology of black-headed ducks is poorly known, and all quantitative information until now stems from Weller's pioneering single-year study more than three decades ago at some of the same sites as those we studied¹⁰. Our study wetlands were within 30 km west or southwest of General Lavalle, Buenos Aires province, Argentina. To detect brood parasitism we conducted systematic surveys of the marshes every two to four days on foot or by canoe. The vegetation was sparse and the large nests were conspicuous enough for us to be confident that we found almost all nests of potential host species breeding on the study area. Nests were identified to species by observing birds on or near nests, or on the basis of distinctive eggs. Parasitism was easily detected because the duck eggs differ markedly from the eggs of all of the major hosts. On each visit, all new eggs were labelled and previous eggs were censused to determine their fate: rejected, depredated, hatched or left over after host chicks hatched. Coots rejected duck eggs mainly by burying them in nesting material, but some were ejected from nests or simply disappeared. Weller¹⁰ also observed rejection by one of the hosts, red-fronted coots.

The two species of coot are by far the most abundant birds in the study marshes, and our exhaustive searches of large tracts of marsh throughout the region did not reveal other suitable hosts that are common but not currently being parasitized. Colonial species such as brown-hooded gulls (*Larus maculipennis*) and white-faced ibis (*Plegadis chichi*) are moderately parasitized where they occur, but colonies are uncommon. Our study corroborates Weller's conclusion that black-headed ducks are obligate brood parasites¹⁰, because our extensive surveys would have discovered duck nests had they been present.

We estimated the relative importance of the two main hosts to overall duckling production by multiplying the total number of duck eggs laid in nests of each host species

by the hatching success for each host, and then calculated the fraction of all ducklings hatching in nests of each host.

Assessing costs to hosts

When assessing whether the presence of duck eggs increased the risk of nest predation for hosts, we excluded parasitized nests in which all duck eggs were rejected. Nests were considered preyed on if all eggs disappeared before they were due to hatch or if we had clear evidence for predation (broken eggs). Clutch size varies considerably between individual hosts, so we assessed hatching success in terms of the number of host eggs that failed to hatch at each nest; this measure includes eggs that disappeared, were rejected or were left over after the rest hatched. Leftover or rejected host eggs were rare, so we primarily measured egg loss. Our more detailed analysis of egg loss in red-fronted coots included experimental nests from which real parasitic eggs were removed quickly after laying ($n = 25$) or in which parasitic duck eggs were experimentally added to unparasitized nests ($n = 22$); these two types of experimental nest enabled us to decouple egg loss due to the act of parasitism itself (damage or removal by parasite) from egg loss due to the presence of duck eggs itself, such as damage to host eggs with subsequent removal by hosts¹⁹. The latter cost favours egg rejection; the former does not.

Mimicry experiments

The white, oval-shaped duck eggs differ from the host eggs in three key visual features—rounder shape, paler background colour and lack of spots (Figs 1h and 2a). We painted domestic chicken eggs and real host eggs to create a series of three egg treatments that increasingly resembled host eggs—the least mimetic ‘white duck’ eggs (experimental versions of real duck eggs) had the wrong shape, background colour and lacked spots, whereas the most mimetic ‘brown coot’ eggs lacked only spots (Fig. 2a). Egg colour and shape vary in real duck eggs (although to a much smaller degree), so these should be feasible evolutionary steps towards mimicry. To avoid a confounding effect of size, we used painted red-gartered coot eggs for the ‘brown coot’ treatment for both hosts, because this species overlaps in size with the duck eggs. For the ‘brown duck’ and ‘white duck’ treatments we used chicken eggs whose length and width both overlapped with those of real duck eggs. We added the experimental eggs to host nests in the laying or early incubation stages and we determined their fates in subsequent visits. Eggs were scored as rejected if found buried in the nest or if observed at least half buried on the final nest visit for nests that hatched or were preyed on before rejection was complete. Non-rejected eggs were scored as accepted only if the nest remained active long enough for rejection to have occurred (at least 10 days for both species).

Intraspecific brood parasitism

In 1997 our studies were conducted primarily on open wetlands where red-fronted coots were absent, so our detailed analysis of intraspecific brood parasitism is restricted to red-gartered coots. Nests were checked every two to four days, which will underestimate parasitism on the basis of unusual egg-laying rates (two or more new eggs per day)²⁸, so we focused on variation in egg features, a reliable method when used conservatively²⁹. Our retrospective assessment of intraspecific brood parasitism from the earlier field seasons, where we did not specifically focus on detecting intraspecific brood parasitism, will greatly underestimate the actual rate of intraspecific brood parasitism: nest checks were relatively infrequent and we would have noticed only the most extreme cases of variation in egg features to detect parasitism^{28–30}.

Received 20 July; accepted 13 September 2004; doi:10.1038/nature03036.

- Rothstein, S. I. A model system for coevolution: avian brood parasitism. *Annu. Rev. Ecol. Syst.* **21**, 481–508 (1990).
- Davies, N. B. *Cuckoos, Cowbirds and Other Cheats* (Poyser, London, 2000).
- Thompson, J. N. The evolution of species interactions. *Science* **284**, 2116–2118 (1999).
- Thompson, J. N., Nuismer, S. L. & Gomulkiwicz, R. Coevolution and maladaptation. *Integ. Comp. Biol.* **42**, 381–387 (2002).
- Brooke, M. de L. & Davies, N. B. Egg mimicry by cuckoos *Cuculus canorus* in relation to discrimination by hosts. *Nature* **335**, 630–632 (1988).
- Marchetti, K., Nakamura, H. & Gibbs, H. L. Host-race formation in the common cuckoo. *Science* **282**, 471–472 (1998).
- Gibbs, H. L., Sorenson, M. D., Marchetti, K., Brooke, M. de L., Davies, N. B. & Nakamura, H. Genetic evidence for female host-specific races in the common cuckoo. *Nature* **407**, 183–186 (2000).
- Payne, R. B. The ecology of brood parasitism in birds. *Annu. Rev. Ecol. Syst.* **8**, 1–28 (1977).
- Lyon, B. E. & Eadie, J. M. Mode of development and interspecific avian brood parasitism. *Behav. Ecol.* **2**, 309–318 (1991).
- Weller, M. W. The breeding biology of the parasitic black-headed duck. *Living Bird* **7**, 169–208 (1968).
- Sorenson, M. D. Effects of intra- and interspecific brood parasitism on a precocial host, the canvasback, *Aythya valisineria*. *Behav. Ecol.* **8**, 153–161 (1997).
- Mason, P. & Rothstein, S. I. Crypsis versus mimicry and the color of shiny cowbird eggs. *Am. Nat.* **130**, 161–167 (1987).
- Roskaft, E., Orians, G. H. & Beletsky, L. D. Why do red-winged blackbirds accept eggs of brown-headed blackbirds? *Evol. Ecol.* **4**, 35–42 (1990).
- Thomson, D. L., Monaghan, P. & Furness, R. W. The demands of incubation and avian clutch size. *Biol. Rev.* **73**, 293–304 (1998).
- Arnold, T. W. *Food Limitation and the Adaptive Significance of Clutch Size in American Coots* (Fulica americana) Thesis, Univ. Western Ontario (1990).
- Freeman, S. Egg variability and conspecific brood parasitism in the *Ploceus* weaverbirds. *Ostrich* **59**, 49–53 (1988).
- Jackson, W. M. in *Parasitic Birds and their Hosts* (eds Rothstein, S. I. & Robinson, S. K.) 407–416 (Oxford University, New York, 1998).
- Lahti, D. C. & Lahti, A. R. How precise is egg discrimination in weaverbirds? *Anim. Behav.* **62**, 1135–1142 (2002).

- Lyon, B. E. Egg recognition and counting reduce costs of avian conspecific brood parasitism. *Nature* **422**, 495–499 (2003).
- Jamieson, I. G., McRae, S. B., Simmons, R. E. & Trewby, M. High rates of conspecific brood parasitism and egg rejection in coots and moorhens in ephemeral wetlands in Namibia. *Auk* **117**, 250–255 (2000).
- Sorenson, M. D. Evidence of conspecific nest parasitism and egg discrimination in the sora. *Condor* **97**, 819–821 (1995).
- Lyon, B. E., Eadie, J. M. & Hamilton, L. D. Parental preference selects for ornamental plumage in American coot chicks. *Nature* **371**, 240–243 (1994).
- Todd, F. S. *Natural History of the Waterfowl* (Ibis Publishing, San Diego, 1996).
- Weller, M. W. Experimental parasitism of American coot nests. *Auk* **88**, 108–115 (1971).
- Lotem, A., Nakamura, H. & Zahavi, A. Rejection of cuckoo eggs in relation to host age: a possible evolutionary equilibrium. *Behav. Ecol.* **3**, 128–132 (1992).
- Lotem, A. Learning to recognize cuckoo nestlings is maladaptive to cuckoo *Cuculus canorus* hosts. *Nature* **362**, 743–745 (1993).
- Davies, N. B., Brooke, M. de L. & Kacelnik, A. Recognition errors and probability of parasitism determines whether reed warblers should accept or reject mimetic cuckoo eggs. *Proc. R. Soc. Lond. B* **263**, 925–931 (1996).
- Frederick, P. C. & Shields, M. A. Corrections for the underestimation of brood parasitism frequency derived from daily nest inspections. *J. Field Ornithol.* **57**, 224–226 (1986).
- Lyon, B. E. Tactics of parasitic American coots: host choice and the pattern of egg dispersion among host nests. *Behav. Ecol. Sociobiol.* **33**, 87–100 (1993).
- McRae, S. B. Identifying eggs of conspecific brood parasites in the field: a cautionary note. *Ibis (Lond. 1859)* **139**, 701–704 (1997).

Acknowledgements We thank C. Benkman, J. Briskie, A. Chaine, N. Davies, M. Hauber, A. Lotem, D. Mock, R. Montgomerie, S. Rothstein, J. Thompson and K. Wasson for comments on the manuscript; the National Geographic Society and the British Broadcasting Corporation (D. Attenborough's *Life of Birds*) for funding; the Flores family, J. Echerran, M. Beade and J. C. Reboreda for logistic support; and V. Meuhter, G. Goggin and A. Carminati for assistance in the field. During part of the study, B.E.L. was supported by the Kananaskis Field Stations of the University of Calgary and University of California, and J.McA.E. by the Dennis G. Raveling Endowment.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.E.L. (lyon@biology.ucsc.edu).

Spatial patterns in species distributions reveal biodiversity change

Robert J. Wilson^{1*}, Chris D. Thomas^{1*}, Richard Fox², David B. Roy³ & William E. Kunin¹

¹Centre for Biodiversity and Conservation, School of Biology, University of Leeds, Leeds LS2 9JT, UK

²Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset BH20 5QP, UK

³NERC Centre for Ecology and Hydrology, Monks Wood, Huntingdon, Cambridgeshire PE28 2LS, UK

* Present addresses: Área de Biodiversidad y Conservación, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, Tulipán s/n, Móstoles, Madrid, E-28933, Spain (R.J.W.); Department of Biology, University of York, P.O. Box 373, York YO10 5YW, UK (C.D.T.).

Interpretation of global biodiversity change is hampered by a lack of information on the historical status of most species in most parts of the world^{1–5}. Here we show that declines and increases can be deduced from current species distributions alone, using spatial patterns of occupancy combined with distribution size. Declining species show sparse, fragmented distributions for their distribution size, reflecting the extinction process; expanding species show denser, more aggregated distributions, reflecting colonization. Past distribution size changes for British butterflies were deduced successfully from current distributions, and former distributions had some power to predict future change. What is more, the relationship between distribution pattern and change in British butterflies

independently predicted distribution change for butterfly species in Flanders, Belgium, and distribution change in British rare plant species is similarly related to spatial distribution pattern. This link between current distribution patterns and processes of distribution change could be used to assess relative levels of threat facing different species, even for regions and taxa lacking detailed historical and ecological information.

Biodiversity assessment relies heavily on combining information on current species status with rates of decline or increase^{1–8}. However, direct evidence of change is unavailable for most taxonomic groups, and for most parts of the world. It is possible to list

species as internationally threatened if they have extremely small distribution sizes or population sizes^{4,6,8}, but rates of decline might better identify levels of threat facing different species, and ignorance of rates of decline might lead to underestimates of extinction risk^{1,3}. Furthermore, species whose distributions have been recorded at finer scales are more likely to meet criteria for listing based on small distribution size^{3,6,9}. Red Data Books are therefore dominated by well-known taxa in well-known regions, even though less well documented groups may be equally endangered and indeed are usually more species-rich than better-known taxa^{2,5}. Here we evaluate whether levels of decline or increase can be deduced from current status alone, because past colonization and extinction processes leave different spatial signatures on species' distribution patterns. Declining species are expected to have sparse distributions because extinctions cause retractions in range to optimal habitats^{10–12} or to locations that have been least affected by wide-acting extinction forces¹³. Increasing species are expected to have more aggregated distributions, resulting from distance-delimited colonization processes¹⁴.

Butterflies form the model system. In Britain, butterfly distributions were mapped comprehensively at 10 km resolution in 1970–82 (ref. 15) and 1995–99 (ref. 16), allowing the measurement of declines and increases¹⁷ and the analysis of spatial distribution patterns. Fine-scale (10-km) distribution records for some species are aggregated in relatively few coarse-scale (100-km) cells (Fig. 1a, b), whereas fine-scale records for other species are scattered sparsely over a wider range of coarse-scale cells (Fig. 1c, d). To quantify this pattern for each species' distribution, we summed first the total area of occupied 10-km cells, and second the total area of occupied 100-km cells, and plotted the logarithm of the area of occupancy (AOO) at each scale against the logarithm of the side of the grid cell (Fig. 1e, f). The slope of this scale–area curve¹⁸ (or range–area relationship¹⁹) gives a measure of the aggregation of each species' distribution (the fractal dimension, D_{ij})^{6,18}. In the most aggregated distributions, a maximum value of $D_{ij} = 2$ indicates that occupied fine-scale cells completely fill each occupied coarse-scale cell. In the sparsest distributions, a minimum value of $D_{ij} = 0$ indicates that each occupied fine-scale cell is located in a separate coarse-scale cell.

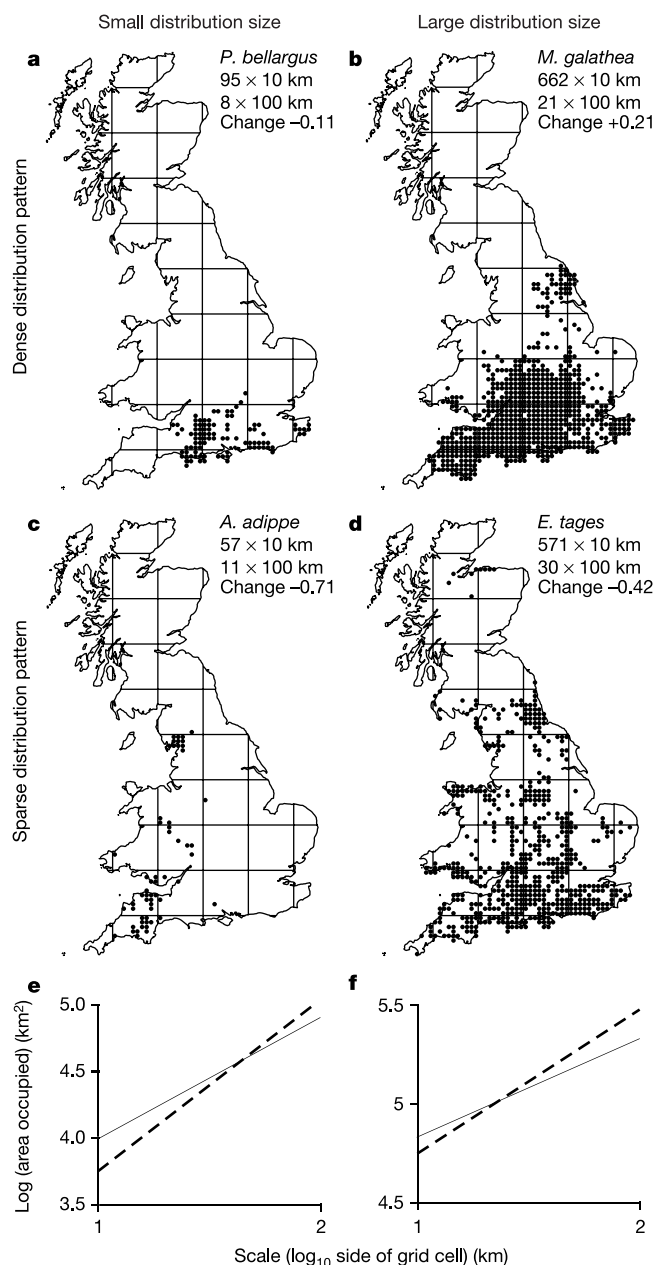


Figure 1 Maps and associated scale–area curves for 1995–99 distributions of British butterfly species. **a–d**, Distribution maps at 10 km resolution, showing the number of 10-km and 100-km cells occupied, and proportional change in 10-km records since 1970–82. **a**, *Polyommatus bellargus*, small, dense distribution; **b**, *Melanargia galathea*, large, dense distribution; **c**, *Argynnis adippe*, small, sparse distribution; **d**, *Erynnis tages*, large, sparse distribution. **e**, **f**, Plots of $\log_{10}(\text{area occupied})$ against scale ($\log_{10}(\text{side of distribution grid cell})$) for *P. bellargus* (solid line) and *A. adippe* (dashed line) (**e**), and *M. galathea* (solid line) and *E. tages* (dashed line) (**f**). Note difference in occupancy (y -axis) values between **e** and **f**. Sparsely distributed species show steeper scale–area curves.

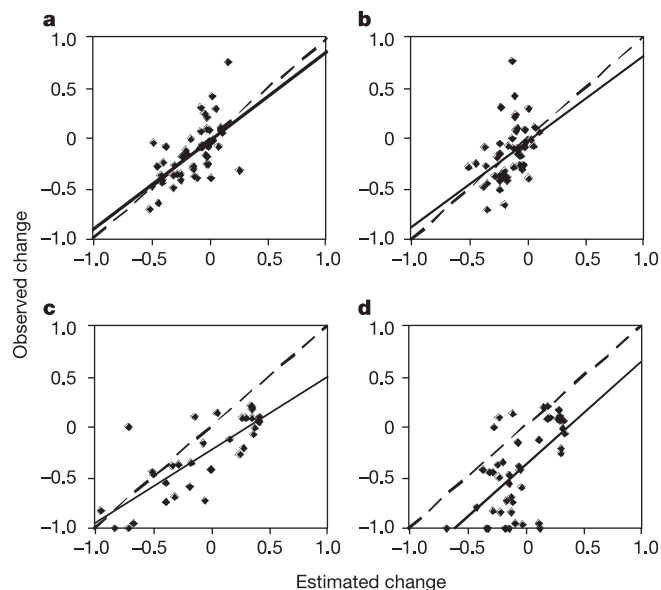


Figure 2 Observed changes in distribution size against changes estimated from distribution pattern and size. **a**, Deductions of past change from 1995–99 distributions in Britain. **b**, Predictions of future change from 1970–82 distributions in Britain. **c**, Deductions of past change from 1991–99 distributions in Flanders. **d**, Predictions of future change from pre-1991 distributions in Flanders. Solid lines are regressions; dashed lines indicate equality of observations and estimations.

Regressions of change in butterfly distribution size (1970–82 to 1995–99) against 1995–99 distribution size at 10 km resolution ($R^2 = 0.16$, $F_{1,49} = 9.53$, $P = 0.003$; change = $0.18\text{AOO} - 0.97$) or fractal dimension ($R^2 = 0.32$, $F_{1,49} = 23.13$, $P < 0.001$; change = $0.48D_{ij} - 0.79$) showed that species with large or aggregated distributions had increased more than species with small or sparse distributions. However, D_{ij} is positively related to AOO ($R^2 = 0.86$, $F_{1,49} = 307.3$, $P < 0.001$), and analysing both variables together showed that species with aggregated distributions for their distribution size had increased the most, whereas species with sparse distributions for their distribution size had declined ($R^2 = 0.43$, $F_{2,48} = 18.06$, $P < 0.001$; change = $0.08 + 1.18D_{ij} - 0.39\text{AOO}$). To determine whether changes in distribution size could be deduced without knowledge of a species' former distribution, simply by examining its current distribution pattern, each species in turn was omitted from the analysis, and the regression of change against D_{ij} and AOO for the remaining species was used to estimate change for the omitted species. These independent deductions of distribution change accounted for 37–38% of variation in observed changes in distribution size (Fig. 2a; $R^2 = 0.38$, $F_{1,49} = 29.77$, $P < 0.001$; observed change = $0.89 (\pm \text{s.e.m. } 0.16) \times \text{estimated change} - 0.02 (\pm \text{s.e.m. } 0.04)$; $R^2 = 0.37$ from phylogenetic generalized least squares (GLS) regression).

To be useful for biodiversity assessment, the relationship of distribution change to spatial distribution pattern should be robust to the nature and quality of distribution data. For British butterflies, the combined ability of distribution pattern and size to deduce distribution change remained strong even if data on change were only available for a small fraction of species (using information from ten randomly selected species to predict change in the remaining 41 species, R^2 ranged from 0.30 to 0.45; Supplementary Information). D_{ij} calculated over a range of scales between 10 and 100 km ranked species consistently (Supplementary Table 3), and the relationship of distribution change with D_{ij} and AOO was consistent, although it was strongest with finer-resolution data (Supplementary Table 4). Neither mobility nor population density was significantly related to the unexplained variation in distribution change (Supplementary Methods), further suggesting that the level of variation explained stems more from data quality than from effects of other explanatory variables.

To test the robustness of the relationship to the method used to quantify aggregation, we also calculated two statistics, D_x (ref. 20) and Ripley's L (ref. 21), on the basis of the number of conspecific records in circles of different-sized radii (10, 20, 50 and 100 km) around each distribution record. In each case, distribution change was positively related to the aggregation of species' distributions relative to their size. The two methods explained similar proportions of variation to that explained by D_{ij} (R^2 for observed change against D_x and AOO ranged from 0.47 to 0.54 at the different scales; for Ripley's L and AOO, R^2 ranged from 0.23 to 0.39) (Supplementary Information).

Species may show time lags in their declines to extinction in fragmented landscapes²² and in their colonization of regions that have become suitable²³, such that past distribution patterns may show the beginnings of range contractions or expansions, and may have some power to predict future change. Distribution change in British butterflies from 1970–82 to 1995–99 was significantly related to fractal dimension and area of occupancy in 1970–82 ($R^2 = 0.23$, $F_{2,48} = 7.32$, $P = 0.002$; change = $0.40 + 1.18D_{ij} - 0.44\text{AOO}$), but 1970–82 distribution pattern and size independently predicted a lower proportion of variation in species' future declines and increases when each species was omitted in turn from the analysis (Fig. 2b; $R^2 = 0.18$, $F_{1,49} = 10.47$, $P = 0.002$; observed change = $0.85 (\pm \text{s.e.m. } 0.26) \times \text{estimated change} - 0.02 (\pm \text{s.e.m. } 0.05)$; phylogenetic GLS $R^2 = 0.16$). The reduced power of the predictions of future change, compared with the deductions of past change, probably result from the earlier survey's less-detailed

distribution data, and because processes of range expansion and contraction themselves lead to pronounced differences in aggregation. Between the two surveys, the distributions of expanding species became more aggregated, and those of contracting species more fragmented, because the residuals for most declining species from the regression of D_{ij} against AOO declined from 1970–82 to 1995–99, whereas the residuals for expanding species increased ($R^2 = 0.10$, $F_{1,49} = 5.70$, $P = 0.02$; distribution change = $-0.15 (\pm \text{s.e.m. } 0.04) + 1.24 (\pm \text{s.e.m. } 0.52) \times \text{change in residuals}$). Processes of distribution change may increase fragmentation in the distributions of already declining species, increasing their risk of future decline by reducing colonization rates and increasing rates of local extinction²⁴. Although predictions of change from the earlier survey were less powerful than deductions from the later survey, the combination of distribution pattern and size still explained two to three times as much variation (16–18%) as the 6–7% explained by distribution size alone ($R^2 = 0.07$, $F_{1,49} = 3.59$, $P = 0.06$; observed change = $0.12\text{AOO} - 0.70$; phylogenetic GLS $R^2 = 0.06$). It therefore does seem possible to predict future dynamics from current species distributions, although additional refinements might be required before such forecasting can be applied directly to conservation.

The potential of this approach for biodiversity assessment is supported by tests of the relationship for another landscape and another taxonomic group. As a test of sensitivity to landscape, the relationship of distribution change with distribution pattern and size in British butterflies was used to deduce changes in the distributions of butterflies in Flanders, Belgium, that have been mapped at 5-km scale in two date categories, pre-1991 and 1991–99 (ref. 25). The British relationship of change with 1995–99 distributions, combined with D_{ij} and AOO for Flanders butterflies in 1991–99, successfully deduced changes in distribution size in Flanders (Fig. 2c; $R^2 = 0.65$, $F_{1,32} = 58.82$, $P < 0.001$; observed change = $0.73 (\pm \text{s.e.m. } 0.10) \times \text{estimated change} - 0.23 (\pm \text{s.e.m. } 0.04)$; phylogenetic GLS $R^2 = 0.64$). Similarly, the British relationship of change with 1970–82 distributions, combined with pre-1991 D_{ij} and AOO for Flanders, significantly predicted future changes in Flanders (Fig. 2d; $R^2 = 0.36$, $F_{1,32} = 18.18$, $P < 0.001$; observed change = $1.00 (\pm \text{s.e.m. } 0.23) \times \text{estimated change} - 0.30 (\pm \text{s.e.m. } 0.05)$; phylogenetic GLS $R^2 = 0.37$) (Supplementary Methods). Flanders declines were worse and increases less pronounced than estimated, possibly because decline is expected to be greater when measured at a finer resolution (5 km in Flanders; 10 km in Britain)^{6,26}. As a test for another taxonomic group, distribution change in rare British plant species was found to be positively related to fractal dimension calculated at scales of 1–10 km or 10–100 km, with a stronger relationship at the finer scale (Supplementary Methods). This suggests that differences in mobility between taxa do not obscure the relationship between aggregation and distribution change; it also supports the notion that the relationship seems stronger when finer-resolution data are used (finer-scale data may be required to detect the relationship in less mobile organisms, for which processes of range change are likely to occur over shorter distances).

Extinction processes typically cause range collapses^{10–13}, leaving small refuge populations that are prone to extinction from both deterministic and stochastic processes²⁷. Consequently, reliable tools are needed to identify range declines and to establish conservation programmes while there remains scope for recovery²⁷. We show that current distribution patterns consistently reveal processes of decline and increase, and that species showing the most fragmented patterns for their distribution size have declined the most. Further work is required to confirm whether the relationship is consistent across landscapes and species assemblages, depending for example on trophic group, habitat and dispersal pattern²⁸, and to assess the coarsest resolution and narrowest geographic extent of distribution data for which the approach is viable. Nevertheless,

analysis of spatial patterns in current species distributions could be a powerful approach for estimating levels of threat for the many regions and taxa with limited information on past distributions and landscape change, or with limited ecological information on species' habitat or climate associations. □

Methods

Distributions

British distribution sizes and fractal dimensions were calculated with the use of all records for each species in England, Wales and Scotland from 1970–82 (ref. 15) and 1995–99 (ref. 16). Occupied 10-km and 100-km squares were calculated with standard Ordnance Survey grid squares. Distribution maps were based on 65,826 species lists (that is, field visits) for 1970–82 (124,978 species $\times$ location distribution records), and 437,690 species lists for 1995–99 (1,548,935 records). Changes in distribution size were calculated by randomly resampling the 1995–99 species lists for each 100-km square to equalize recorder effort between the two periods¹⁷. Proportional change in distribution size was the number of 10-km squares occupied in 1970–82 subtracted from the number of squares occupied in the resampled 1995–99 data, divided by the number of squares occupied in 1970–82.

The analysis includes all resident butterfly species that have regularly been observed in Britain since 1970, apart from those species for which more than 40% of occupied grid squares in either period were migrants, vagrants or deliberate introductions¹⁶. Fifty-one species were included, none of which occupied 10% or more of their 100-km squares through migrants or introductions¹⁶ (Supplementary Methods).

Area of occupancy and fractal dimension

Area of occupancy (in km²) of each British species in each period was calculated at two scales, first by summing the areas of occupied 10-km squares, and second those of occupied 100-km squares. Log₁₀AOO at each scale was plotted against log₁₀(side of grid square (in km)), and the fractal dimension (D_{ij}) was the slope of this scale-area curve, subtracted from 2 (ref. 18). The fractal is used in these analyses as a descriptive measurement of spatial aggregation over a narrow range of scales: we do not imply that these species have 'truly' fractal distributions over multiple scales.

Estimating change

To estimate distribution change independently for each British species, 51 linear regressions of distribution change against D_{ij} and AOO (calculated at 10-km scale) were performed, leaving out each species in turn. Coefficients from the analysis with the remaining 50 species were used to estimate change with the use of D_{ij} and AOO for each omitted species. We tested the predictions by using a linear regression of observed change against estimated change for all 51 species (Fig. 2a, b).

Phylogenetic regression

The main results presented refer to linear regressions with species as independent data points, on the assumption that information on phylogenetic relatedness was not available (as might be typical for poorly recorded taxa). We tested the sensitivity of the results to phylogenetic relatedness with GLS regressions²⁹ implemented in the software package COMPARE³⁰ (Supplementary Methods). These analyses suggested that the evolutionary constraint acting on distribution pattern, size and change was small, and gave results consistent with those using species as independent data points. For analyses in which the proportion of variation explained is most important, R^2 from the phylogenetic GLS regression is presented alongside the raw results in the main text. Results from all other phylogenetic regressions are presented in Supplementary Information.

Flanders butterflies and British plants

Details of analyses are presented in Supplementary Methods.

Received 23 July; accepted 17 September 2004; doi:10.1038/nature03031.

1. Akçakaya, H. R. *et al.* Making consistent IUCN classifications under uncertainty. *Conserv. Biol.* **14**, 1001–1013 (2000).
2. Groombridge, B. & Jenkins, M. D. *World Atlas of Biodiversity* (Univ. California Press, Berkeley, CA, 2002).
3. Keith, D. A., Auld, T. D., Ooi, M. K. J. & Mackenzie, D. E. Sensitivity analyses of decision rules in World Conservation Union (IUCN) Red List criteria using Australian plants. *Biol. Conserv.* **94**, 311–319 (2000).
4. Mace, G. M. & Lande, R. Assessing extinction threats: towards a re-evaluation of IUCN threatened species categories. *Conserv. Biol.* **5**, 148–157 (1991).
5. World Conservation Monitoring Centre. *Global Biodiversity: Status of the Earth's Living Resources* (Chapman & Hall, London, 1992).
6. Hartley, S. & Kunin, W. E. Scale dependency of rarity, extinction risk, and conservation priority. *Conserv. Biol.* **17**, 1559–1570 (2003).
7. Warren, M. S., Barnett, L. K., Gibbons, D. W. & Avery, M. I. Assessing national conservation priorities: an improved red list of British butterflies. *Biol. Conserv.* **82**, 317–328 (1997).
8. World Conservation Union (IUCN). *IUCN Red List Categories. Version 3.1.* (World Conservation Union, Gland, Switzerland, 2001).
9. Cowley, M. J. R., Thomas, C. D., Thomas, J. A. & Warren, M. S. Flight areas of British butterflies: assessing species status and decline. *Proc. R. Soc. Lond. B* **266**, 1587–1592 (1999).
10. Johnson, C. N. Species extinction and the relationships between distribution and abundance. *Nature* **394**, 272–274 (1998).
11. Brown, J. H., Mehlman, D. W. & Stevens, G. C. Spatial variation in abundance. *Ecology* **76**, 2028–2043 (1995).
12. Rodriguez, A. & Delibes, M. Internal structure and patterns of contraction in the geographic range of the Iberian Lynx. *Ecography* **25**, 314–328 (2002).

13. Channell, R. & Lomolino, M. V. Dynamic biogeography and conservation of endangered species. *Nature* **403**, 84–86 (2000).
14. Shigesada, N. & Kawasaki, K. *Biological Invasions: Theory and Practice* (Oxford Univ. Press, Oxford, UK, 1997).
15. Heath, J., Pollard, E. & Thomas, J. A. *Atlas of Butterflies in Britain and Ireland* (Viking Books, Harmondsworth, UK, 1984).
16. Asher, J. *et al.* *The Millennium Atlas of Butterflies in Britain and Ireland* (Oxford Univ. Press, Oxford, UK, 2001).
17. Warren, M. S. *et al.* Rapid responses of British butterflies to opposing forces of climate and habitat change. *Nature* **414**, 65–69 (2001).
18. Kunin, W. E. Extrapolating species abundance across spatial scales. *Science* **281**, 1513–1515 (1998).
19. Ostling, A., Harte, J. & Green, J. L. Self-similarity and clustering in the spatial distribution of species. *Science* **290** (Suppl.), 671a (2000).
20. Condit, R. *et al.* Spatial patterns in the distribution of tropical tree species. *Science* **288**, 1414–1418 (2000).
21. Ripley, B. D. *Spatial Statistics* (John Wiley & Sons, New York, 1981).
22. Brooks, T. M., Pimm, S. L. & Oyugi, J. O. Time lag between deforestation and bird extinction in tropical forest fragments. *Conserv. Biol.* **13**, 1140–1150 (1999).
23. Hill, J. K., Thomas, C. D. & Huntley, B. Climate and habitat availability determine 20th century changes in a butterfly's range margins. *Proc. R. Soc. Lond. B* **266**, 1197–1206 (1999).
24. Hanski, I. *Metapopulation Ecology* (Oxford Univ. Press, Oxford, 1999).
25. Maes, D. & Van Dyck, H. Butterfly diversity loss in Flanders (north Belgium): Europe's worst case scenario? *Biol. Conserv.* **99**, 263–276 (2001).
26. Thomas, C. D. & Abery, J. C. G. Estimating rates of butterfly decline from distribution maps: the effect of scale. *Biol. Conserv.* **73**, 59–65 (1995).
27. Caughley, G. Directions in conservation biology. *J. Anim. Ecol.* **63**, 215–244 (1994).
28. Linder, E. T., Villard, M.-A., Maurer, B. A. & Schmidt, E. V. Geographic range structure in North American landbirds: variation with migratory strategy, trophic level, and breeding habitat. *Ecography* **23**, 678–686 (2000).
29. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667 (1997).
30. Martins, E. P. COMPARE, version 4.6. *Computer programs for the statistical analysis of comparative data* <http://compare.bio.indiana.edu/> (Department of Biology, Indiana University, Bloomington IN, 2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the contributors to the butterfly and plant distribution surveys in Britain and Flanders; J. Asher and J. A. Thomas for calculating British rates of decline; and M. de la Cruz, A. Escudero, S. Hartley, J. Perry and M. Pocock for assistance with analyses. The work was supported by the UK Natural Environment Research Council. Maps were produced in DMap.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.J.W. (rwilson@escet.urjc.es).

Identification of human brain tumour initiating cells

Sheila K. Singh^{1,2,3}, Cynthia Hawkins^{1,4}, Ian D. Clarke^{1,2}, Jeremy A. Squire⁶, Jane Bayani⁶, Takuichiro Hide^{1,2}, R. Mark Henkelman⁵, Michael D. Cusimano^{3,7} & Peter B. Dirks^{1,2,3}

¹The Arthur and Sonia Labatt Brain Tumor Research Centre, ²Program in Developmental Biology, ³Division of Neurosurgery, ⁴Department of Pediatric Laboratory Medicine, and ⁵Integrative Biology Program, The Hospital for Sick Children and University of Toronto, 555 University Avenue, Toronto, M5G 1X8, Canada

⁶Ontario Cancer Institute and University of Toronto, 610 University Avenue, Toronto, M5G 2M9, Canada

⁷Division of Neurosurgery, St Michael's Hospital and University of Toronto, 30 Bond Street, Toronto M5B 1W8, Canada

The cancer stem cell (CSC) hypothesis suggests that neoplastic clones are maintained exclusively by a rare fraction of cells with stem cell properties^{1,2}. Although the existence of CSCs in human leukaemia is established^{3,4}, little evidence exists for CSCs in solid tumours, except for breast cancer⁵. Recently, we prospectively isolated a CD133⁺ cell subpopulation from human brain tumours that exhibited stem cell properties *in vitro*⁶. However,

the true measures of CSCs are their capacity for self renewal and exact recapitulation of the original tumour^{1,2,7}. Here we report the development of a xenograft assay that identified human brain tumour initiating cells that initiate tumours *in vivo*. Only the CD133⁺ brain tumour fraction contains cells that are capable of tumour initiation in NOD-SCID (non-obese diabetic, severe combined immunodeficient) mouse brains. Injection of as few as 100 CD133⁺ cells produced a tumour that could be serially transplanted and was a phenocopy of the patient's original tumour, whereas injection of 10⁵ CD133⁻ cells engrafted but did not cause a tumour. Thus, the identification of brain tumour initiating cells provides insights into human brain tumour pathogenesis, giving strong support for the CSC hypothesis as the basis for many solid tumours⁵, and establishes a previously unidentified cellular target for more effective cancer therapies.

Several studies have identified stem-like cells in brain tumour cultures^{8–10}. By cell sorting for the cell surface antigen CD133 (refs 11–13), we demonstrated that a functional hierarchy exists in the brain tumour cell population *in vitro*⁶. To determine whether the CD133⁺ human brain tumour cells were capable of tumour initiation *in vivo*, we compared the abilities of CD133⁺ versus CD133⁻ tumour cells to initiate tumour formation in NOD-SCID mouse brains. Immediately after surgical resection, solid, non-metastatic, brain tumour masses (medulloblastomas from three children, glioblastomas from three adults and one childhood glioblastoma; see Supplementary Table 1) were dissociated into single-cell suspensions. *In vitro* primary sphere formation assays were performed for all uncultured tumours and flow cytometric quantification of CD133 expression was performed on seven acutely dissociated tumours and two corresponding tumour xenografts before serial retransplantation. The CD133⁺ fraction among highly aggressive glioblastomas (GBMs) ranged from 19 to 29%, and among medulloblastomas ranged from 6 to 21%, and correlated closely with an *in vitro* primary sphere formation assay (which we used to quantify stem cell frequency; see Supplementary Table 2). After dissociation, magnetic bead cell sorting was used to separate the CD133⁺ brain tumour cells from their CD133⁻ counterparts. Sorted cells were transplanted into the frontal lobe of six-week-old NOD-SCID mice.

Most intracranial xenograft models using human tumour cell lines or primary cultures require 10⁵ to 10⁶ cells for tumour engraftment and formation^{14,15}. Analysis of mouse brains following CD133⁺ engraftment revealed that as few as 100 CD133⁺ cells were sufficient for the formation of human brain tumours in NOD-SCID mice that were analysed at 12–24 weeks post-injection (Fig. 1a–c) (see Supplementary Table 3). Injection of 50,000 to 100,000 of CD133⁻ cells did not form tumours in the NOD-SCID brains, and histological examination at 12 weeks revealed only a glial scar tract from initial injection (Fig. 1d). A total of 15 CD133⁻ injections and 19 CD133⁺ injections were performed. Of the 19 mice injected with CD133⁺ medulloblastoma or glioblastoma cells, 16 developed brain tumours (see Supplementary Table 3).

CD133⁺ xenografts from two classic medulloblastomas showed small round blue cell morphology and demonstrated characteristic histologic structures (Homer–Wright rosettes)¹⁶ (Fig. 2). CD133⁺ xenografts from a third desmoplastic medulloblastoma recapitulated the more primitive cytoarchitecture associated with this subtype (Supplementary Fig. 1).

To further characterize the tumour phenotype of the CD133⁺ medulloblastoma xenografts, we performed detailed immunohistochemical analyses of each transplanted tumour and of the original patients' tumours that were used for initial histologic diagnosis (see Supplementary Table 3). The CD133⁺ xenografts from all three medulloblastomas expressed the cytoplasmic primitive intermediate filaments nestin¹⁷ (Fig. 2) and vimentin (data not shown), frequently used as neural precursor cell markers, although they are not definitive neural stem cell markers. All three medulloblastoma

xenografts expressed the neuronal marker β III-tubulin in a significant number of tumour cells, as did the original tumours from the patients (Fig. 2). Expression of the astrocyte marker GFAP (glial fibrillary acidic protein) was also seen in a minority of the human tumour cells in two of three medulloblastoma xenografts (Fig. 2). The GFAP-positive cells are morphologically neoplastic cells and did not demonstrate hybridization with mouse centromeric probes by interphase fluorescence *in situ* hybridization (FISH, data not shown) supporting that they were derived from the human cell transplant and were not trapped normal mouse astrocytes. Proliferative indices, indicated by MIB-1 (Ki67) immunostaining, were comparable in the xenografts and the original tumours (Fig. 2).

CD133⁺ xenografts from the adult and childhood GBMs demonstrated classical histopathological features of this tumour type. Both the classical GBM and the diffusely infiltrative GBM (Fig. 3 and Supplementary Fig. 1) possessed the four diagnostic criteria as defined by the World Health Organization¹⁸. In each case, the phenotype of the mouse xenograft matched that of the patient's original tumour.

CD133⁺ GBM xenograft phenotype was further studied by immunohistochemistry (Fig. 3). GBM xenografts also expressed nestin, showing a heterogeneous staining pattern that was consistent with the patients' original tumours. MIB-1 immunostaining showed a high degree of proliferation in both xenografts and the original tumours. CD133⁺ GBM xenografts also demonstrated characteristic expression of GFAP resembling the original patients' tumours, but also showed immunostaining for the neuronal marker MAP2 (microtubule associated protein) in the human tumour cells, which did not label with mouse-specific centromeric probes by FISH (data not shown). This finding of the expression of markers for both neuronal and astrocyte lineages within the CD133⁺ xenograft reflects an ability of the brain tumour initiating cells (BTIC) for multilineage, albeit cancerous, differentiation *in vivo*. In addition, immunostaining for p53 also revealed a similar increase in nuclear staining pattern for this tumour suppressor protein in the patient tumour and xenograft.

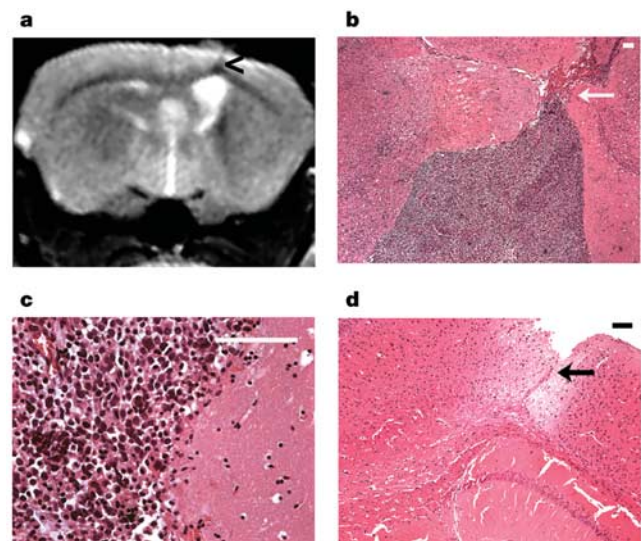


Figure 1 CD133⁺ tumour cells initiate tumours upon intracranial transplantation into the adult NOD-SCID mouse forebrain. **a**, Magnetic resonance imaging (MRI) scan of a mouse injected with 1,000 CD133⁺ medulloblastoma cells shows an enhancing mass under the injection tract (arrowheads) 14 weeks post-injection. **b**, **c**, Low (**b**) and high (**c**) magnification histological sections of the xenograft show a highly cellular mass below the injection site (white arrow in **b**). **d**, Histological section of mouse brain injected with CD133⁻ medulloblastoma cells shows the injection tract (black arrow), but no tumour formation. Scale bar on all panels represents 100 microns.

CD133 immunostaining of the GBM xenografts revealed islands of positive cells, or single positive cells, amid large groups of negative cells; this indicates that not every cell in the xenograft is CD133⁺ (Fig. 4a). Double immunostaining for CD133 and GFAP in CD133⁺ xenografts from GBM samples demonstrated that these markers were expressed in different tumour cell subpopulations, suggesting that undifferentiated and differentiated tumour cells coexist within the transplanted tumours (Fig. 4b). A purified but small number of transplanted CD133⁺ cells (Fig. 4c) resulted in a heterogeneous primary xenograft consisting of a minority of CD133⁺ cells (19–22%) and a majority of CD133[−] cells (78–81%) (Fig. 4d). These results suggest that a tumour hierarchy exists in which the CD133⁺ cells may generate CD133[−] tumour cells. These data are consistent with our previous observations, which demonstrate that only the CD133⁺ fraction is proliferating and

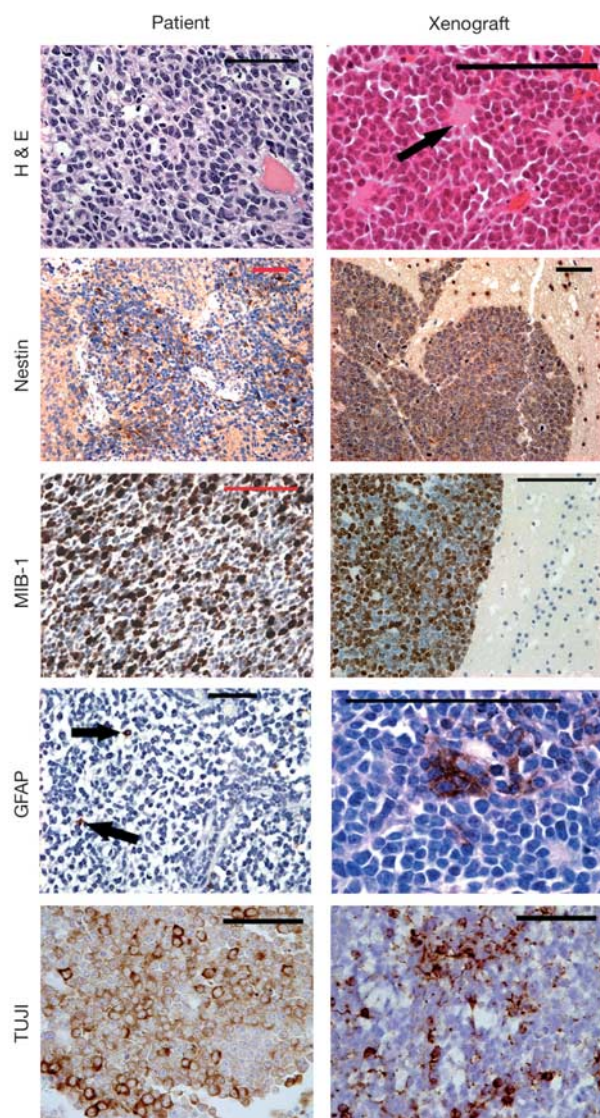


Figure 2 CD133⁺ xenograft from a medulloblastoma (right side) resembles the original patient tumour (left side). H&E of a CD133⁺ xenograft from patient 3 shows classical medulloblastoma cytoarchitecture, resembling the original patient's histology; a Homer–Wright rosette is indicated (arrow). The xenograft and the original tumour both express the neural precursor cell marker nestin (brown) and the neuronal marker β III-tubulin (TUJ1), and show a high proliferative index (MIB-1, brown nuclear staining), which is further increased in the xenograft. The astrocyte cell marker GFAP is also expressed in a small number of cells in the patient tumour and xenograft (brown, see arrows). Scale bar, 100 microns.

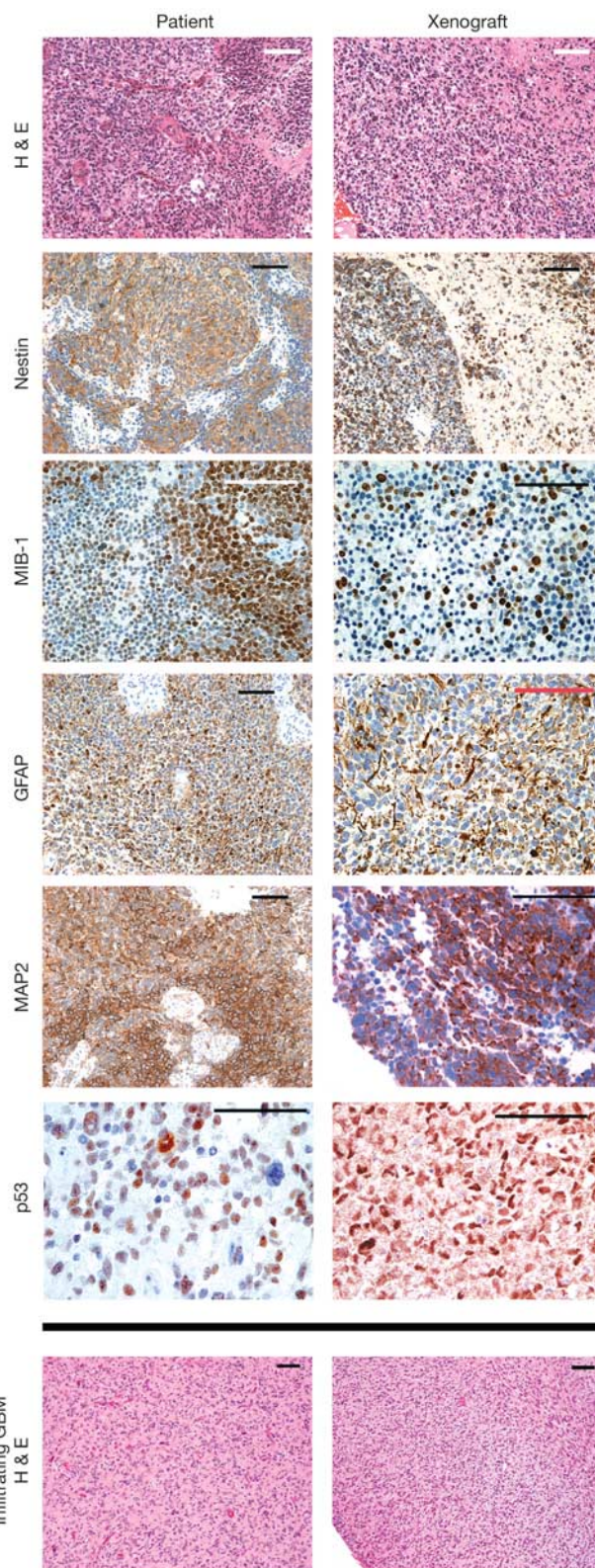


Figure 3 CD133⁺ xenograft from a GBM (right) resembles the original patient tumour (left). H&E section of a CD133⁺ xenograft from patient 5 shows histological features of a GBM, reflecting the patient's original tumour histology. Both the CD133⁺ BTIC xenograft and the patient's original tumour show expression of the neural precursor marker nestin, high proliferative indices (MIB-1), p53 expression, and expression of both neuronal and astrocyte differentiated cell markers (MAP2 and GFAP). There are many invading human-specific nestin-positive tumour cells beyond the tumour borders in the xenograft. A diffusely infiltrative GBM (patient 7) is also depicted by H&E (bottom). Scale bar, 100 microns.

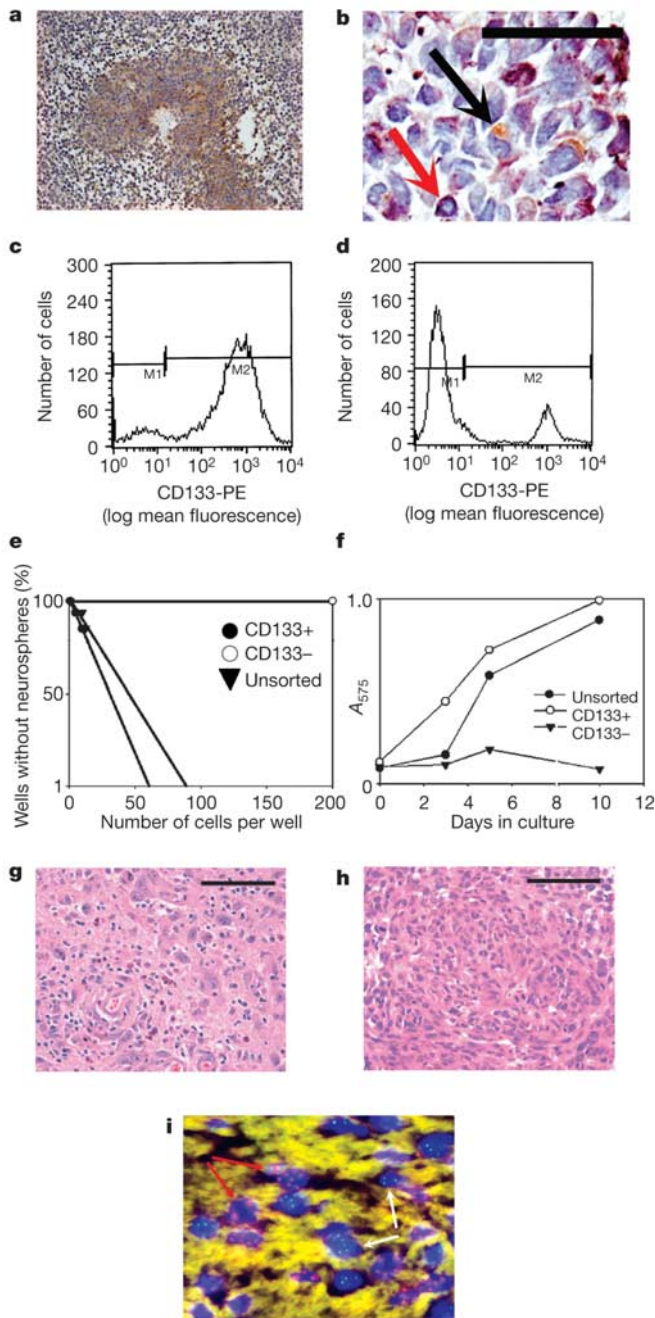


Figure 4 Only the CD133⁺ BTIC, and not the CD133⁻ tumour cell, is capable of tumour initiation. **a**, A subpopulation of cells from a CD133⁺ GBM xenograft are CD133⁺ (brown). **b**, Undifferentiated CD133⁺ (brown, black arrow) and differentiated GFAP⁺ (purple, red arrow) cells coexist in distinct subpopulations in this GBM xenograft. **c**, Purity of CD133⁺ cells (84%) before initial xenotransplantation. M1, CD133⁻ cells; M2, CD133⁺ cells. **d**, Example of acutely dissociated primary xenograft; 19% CD133⁺, 81% CD133⁻. **e, f**, Limiting dilution analysis (**e**) and proliferation assays (**f**) of the GBM sorted for CD133 show that CD133⁺ cells exclusively proliferate and self renew *in vitro*. Absorbance in **f** was measured at 575 nm. **g, h**, H&E sections of a recurrent paediatric GBM from the original patient (**g**, patient 6) and from a secondary xenograft from this tumour (**h**). **i**, Paraffin FISH of CD133⁻ GBM xenograft shows that CD133⁻ human tumour cells survive after xenotransplant but do not form tumours (human pan-centromeric probes in green, white arrows; mouse pan-centromeric probe in red, red arrows). Scale bar, 100 microns.

demonstrating self renewal *in vitro* (Fig. 4e, f).

A key property of all normal and cancer stem cells is the ability to self renew, a feature that can only be tested by serial passage. To determine whether the CD133⁺ BTIC had self-renewal capacity, we performed serial retransplantation experiments from primary xenografts derived from a paediatric and an adult GBM (patients 6 and 7) that grew from an initial injection of 1,000 CD133⁺ cells. After 6 weeks, the primary tumour was excised from the mouse brain and 1,000 CD133⁺ cells were isolated and then reinjected into secondary mice. After 5 weeks, we observed that two out of two secondarily xenografted mice from the paediatric GBM and three of three mice from the adult GBM had brain tumours that recapitulated the phenotype of the original patient tumour and the primary xenograft (Fig. 4g, h), providing direct evidence for the self-renewal capacity of this population.

Most notably, none of the mice injected with CD133⁻ tumour cells, purified to near homogeneity, developed brain tumours when analysed at 12 weeks post-injection. On the basis of analysis of flow histograms of tumour cells sorted for CD133, the contaminating CD133⁺ cells in the purified CD133⁻ population represent a weakly staining population that has very low CD133 protein expression, rather than the highly staining true CD133⁺ sorted population (see Methods). To determine whether viable CD133⁻ human cells had engrafted into the mouse brain but did not form tumours, we performed interphase FISH analysis of murine brain sections using species-specific centromeric probes to detect the presence of human cells at the injection site (Fig. 4i). For each of the CD133⁻ injections from five medulloblastoma and glioblastoma human brain tumour samples (patients 1–5), human cells could be found in small clusters near the original injection site, but these cells did not form a nodule or mass; this indicates that their inability to form tumours is cell intrinsic, and not owing to an inability to be adequately supported in the brain environment following transplant.

To demonstrate that all engrafted cells have undergone a transformation event and do not represent normal brain cells, we conducted molecular cytogenetic analysis by spectral karyotyping (SKY) and interphase FISH using preparations obtained directly from briefly cultured tumour cells sorted for CD133. We found that both CD133⁻ and CD133⁺ tumour cells from a medulloblastoma exhibited chromosomal instability for chromosome 17. The normal signal pattern of two centromere signals and two TP53 signals was detected in less than 4% of sorted cells that were subsequently xenotransplanted, indicating a low rate of potential normal cell contamination (see Supplementary Fig. 2). SKY analysis of GBM specimens bore chromosomal aberrations characteristic of this tumour type (see Supplementary Fig. 3). Further interphase FISH analysis of this GBM using centromere 7 indicated that nearly 80% of both CD133⁺ and CD133⁻ cells exhibited an abnormal karyotype, consistent with the SKY findings. Paraffin FISH studies performed on CD133⁺ xenografts from an adult GBM also indicated that the majority of cells bore evidence of transformation, with amplification of the EGFR gene identical to that of the patient's tumour (see Supplementary Fig. 4). These cytogenetic data indicate that the transplanted cells, whether CD133⁻ or CD133⁺, have an abnormal karyotype and aneusomy pattern inconsistent with significant contamination by normal cells. The CD133⁺ and CD133⁻ cells have the same cytogenetic alterations, suggesting that they are clonally derived.

Together, these data indicate that the CD133⁺ human brain tumour cell fraction from tumours of different types, from both adults and children, contain brain tumour initiating cells that exclusively initiate tumour formation in immunodeficient mice. BTIC have potent *in vivo* self-renewal and proliferative capacities, generating tumours that are a phenocopy of the patient's tumour following engraftment with as few as 100 CD133⁺ cells. Thus, the BTIC possess all the key properties ascribed to a stem cell. These results indicate that brain cancer should not be viewed as blocked or

frozen differentiation; rather, the brain tumour initiating cell clone exhibits patterns of abnormal differentiation. Although our model demonstrates that tumours are initiated by CD133⁺ cells with stem cell properties, whether the cancer-initiating event occurs in a normal stem cell or in normal progenitor or differentiated cells that have reacquired stem cell properties remains to be determined.

Although the frequency of cancer-initiating cells in brain tumours is higher than reported for other cancers^{4,5}, we suggest that the cancer-initiating frequency may be higher because these tumours are among the most aggressive known human cancers and our patient samples were also derived from within the patient's untreated primary tumour masses. Further *in vivo* limiting dilution analysis will continue to more accurately determine the CSC frequency within the CD133⁺ population.

These results suggest that brain cancers are hierarchically organized and lend strong support for the CSC hypothesis, extending this idea beyond the original studies on human leukaemias^{3,4,7,19}. The CSC hierarchy may be functionally elucidated as more surface markers for neural stem cells emerge, and more potent stem cell activity may be found in a CD133⁺ subpopulation. Leukaemias have been suggested to originate in stem cell⁴ as well as progenitor populations^{20,21}, and the question of cell of origin remains unclear in brain tumours^{22–25}.

The identification of an *in vivo* tumour-initiating cell from human brain tumours of different phenotypes provides a powerful tool to investigate the tumorigenic process in the central nervous system. We have also efficiently created a model that can be used to study each individual patient's brain tumour. The cellular heterogeneity of brain tumours dictates that if only a small fraction of cancer stem cells is capable of regenerating the tumour, bulk therapy may fail to target the tumour-initiating cell, allowing for disease progression or relapse. The functional analysis of the BTIC may give new insight into patient prognosis that may then warrant individual tailoring of therapy.

Note added in proof: While our manuscript was under consideration, Galli *et al.*²⁸ demonstrated that glioblastoma cell lines, established by culture in neurosphere conditions, could proliferate, self renew and differentiate into multiple lineages *in vitro*. Cerebral injection of 200,000 of these tumour sphere cells could also generate tumours *in vivo*, and after repeat culture, could initiate phenotypically similar tumours, in a secondary mouse. □

Methods

Primary tumour sphere culture

Tumour samples were obtained from consenting patients, as approved by the Research Ethics Boards at The Hospital for Sick Children and St Michael's Hospital (Toronto). Tumours were washed, acutely dissociated in oxygenated artificial cerebrospinal fluid (CSF) and subject to enzymatic dissociation. Tumour cells were then briefly placed in tumour sphere media (TSM) to allow for recovery following enzymatic dissociation⁶. On average, each tumour specimen yielded two to five million cells.

Magnetic cell sorting and flow cytometry

Cells were labelled with 1 µL CD133/1 microbeads per 1 million cells using the Miltenyi Biotec CD133 cell isolation kit, as previously described⁶, between 1–24 h post-dissociation (median 2.75 h). Tumour cells were generally very adherent and required frequent mechanical and chemical trituration to prevent cell clumping during sorting. Aliquots of CD133⁺ and CD133[−] sorted cells were evaluated for purity by flow cytometry with a FACSCalibur machine (BD Biosciences), using CD133/2 (293C3)-PE antibody (Miltenyi Biotec). Purities ranged from 70 to 91% for CD133⁺ cells (median 84%), and 87.5 to 99.5% (median 99.5%) for CD133[−] cells. The small impurity of CD133⁺ cells in the CD133[−] fraction quantified by FACS showed very low mean fluorescence in comparison to sorted CD133⁺ cells, which had high fluorescence; this suggests that these contaminating CD133⁺ cells in the CD133[−] fraction are phenotypically different from the sorted CD133⁺ population that were used in our transplantation studies.

Intracranial cell transplantation into NOD-SCID mice

Within 1 to 16 h of magnetic bead cell sorting, purified populations of CD133⁺ and CD133[−] cells were resuspended in 10 µL of phosphate buffered saline (PBS), in aliquots of 50,000, 10,000, 5,000 and 1,000 CD133⁺ cells (as few as 100 CD133⁺ cells were used for patients 6 and 7) and 50,000 or 100,000 CD133[−] cells. These aliquots were injected

stereotactically into 6- to 8-week-old NOD-SCID mouse frontal cortex, following administration of general anaesthesia. The injection coordinates were 3 mm to the right of the midline, 2 mm anterior to the coronal suture and 3 mm deep.

Mouse brain fixation and histopathology

Mice were killed and their brains were immediately removed and fixed in 4% paraformaldehyde for 24 h, then transferred to 70% ethanol. Mouse brains were processed on a Tissue-Tek VIP (Sakura) and embedded in paraffin. Brains were sectioned at 5-µm thickness on a Microm HM 200 cryotome (Eryostar), and stained with haematoxylin-and-eosin (H&E) as per standard histopathological technique.

Immunohistochemistry

Paraffin embedded, 5-µm formalin fixed tissue sections were mounted on microscope slides. Tissue sections were then dried overnight at 60 °C, dewaxed in xylene and rehydrated with distilled water. With the exception of sections stained for GFAP, all sections were treated with heat-induced epitope retrieval technique (HIER) using a citrate buffer at pH 6.0. Incubation with the following antibodies was performed for 1 h at room temperature: βIII-tubulin (Chemicon, 1:500), GFAP (DakoCytomation, 1:1,000), MIB-1 (DakoCytomation, 1:20) nestin (Chemicon, 1:200), MAP2 (Sigma, 1:1,000) and p53 (DAKO, 1:20). Immunostaining was performed on a Ventana NEXES auto-immunostainer (Ventana Medical Systems). Double immunostaining for CD133 and GFAP used anti-human CD133 (Miltenyi Biotec, 1:5) incubated overnight at room temperature, followed by anti-human GFAP (DAKO, 1:1,000) for 1 h at room temperature. Human placenta^{11,26} and human retinoblastoma specimens were used as positive controls for CD133. Immunodetection was performed using the Elite Vector Stain ABC System (Vector Laboratories). The counterstain of preference was haematoxylin for nuclear detail.

In vitro limiting dilution analysis (primary sphere formation assays)

Limiting dilution assay on tumour cells that had been sorted for CD133 was performed and 0.37 intercepts calculated⁶. Primary sphere formation assays were performed on the entire acutely dissociated tumour cell population on day 0 to quantify stem cell frequency within the tumour, as previously described⁶. Cell proliferation assays were performed on days 0, 3, 5 and 7 post-plating using the Roche MTT-based colorimetric assay cell proliferation kit 1. Cells were plated in 96-well microwell plates in 0.1-ml volumes of TSM, at a density of 1,000 cells per well.

Cytogenetic preparation, FISH analysis and spectral karyotyping of tumour sphere cells

Short-term cultured sorted or unsorted cells were colcemid treated, hypotonically swelled and methanol:acetic acid treated as previously described²⁷. Interphase FISH analysis was performed using commercially available probes for centromere 17 and the locus specific probe for p53, as well as the centromere 7 and EGFR probe set (Vysis), and used according to the manufacturer's instructions. Pancentromeric mouse and human specific probes (Cedarlane) were also used. The slides were visualized and imaged using the Quips Imaging System (Vysis). Two-hundred nuclei were scored for each cell population and tabulated. Spectral karyotyping was performed on metaphase spreads using the commercially available probes supplied by Applied Spectral Imaging, and analysed as previously described²⁷.

Received 7 September; accepted 22 October 2004; doi:10.1038/nature03128.

- Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
- Pardal, R., Clarke, M. & Morrison, S. Applying the principles of stem-cell biology to cancer. *Nat. Rev. Cancer* **3**, 895–902 (2003).
- Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
- Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Med.* **3**, 730–737 (1997).
- Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
- Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63**, 5821–5828 (2003).
- Dick, J. E. Breast cancer stem cells revealed. *Proc. Natl Acad. Sci. USA* **100**, 3547–3549 (2003).
- Ignatova, T. N. *et al.* Human cortical glial tumors contain neural stem-like cells expressing astroglial and neuronal markers *in vitro*. *Glia* **39**, 193–206 (2002).
- Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumors. *Proc. Natl Acad. Sci. USA* **100**, 15178–15183 (2003).
- Kondo, T., Setoguchi, T. & Taga, T. Persistence of a small subpopulation of cancer stem-like cells in the C6 glioma cell line. *Proc. Natl Acad. Sci. USA* **101**, 781–786 (2004).
- Miraglia, S. *et al.* A novel five-transmembrane hematopoietic stem cell antigen: isolation, characterization, and molecular cloning. *Blood* **90**, 5013–5021 (1997).
- Corbeil, D., Roper, K., Weigmann, A. & Huttner, W. B. AC133 hematopoietic stem cell antigen: human homologue of mouse kidney prominin or distinct member of a novel protein family? *Blood* **91**, 2625–2626 (1998).
- Uchida, N. *et al.* Direct isolation of human central nervous system stem cells. *Proc. Natl Acad. Sci. USA* **97**, 14720–14725 (2000).
- Houchens, D. P., Ovejera, A. A., Riblet, S. M. & Slagel, D. E. Human brain tumor xenografts in nude mice as a chemotherapy model. *Eur. J. Cancer Clin. Oncol.* **19**, 799–805 (1983).
- Hu, B. *et al.* Angiopoietin-2 induces human glioma invasion through the activation of matrix metalloproteinase-2. *Proc. Natl Acad. Sci. USA* **100**, 8904–8909 (2003).
- Russell, D. & Rubenstein, L. *Pathology of Tumors of the Central Nervous System* (Williams and Wilkins, Baltimore, 1989).

17. Lendahl, U., Zimmerman, L. B. & McKay, R. D. CNS stem cells express a new class of intermediate filament protein. *Cell* **60**, 585–595 (1990).
18. Kleihues, P. *et al.* The WHO classification of tumors of the nervous system. *J. Neuropathol. Exp. Neurol.* **61**, 215–225; discussion 226–229 (2002).
19. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature Immunol.* **5**, 738–743 (2004).
20. Cozzio, A. *et al.* Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. *Genes Dev.* **17**, 3029–3035 (2003).
21. Passegue, E., Jamieson, C. H., Ailles, L. E. & Weissman, I. L. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? *Proc. Natl Acad. Sci. USA* **100**(suppl. 1), 11842–11849 (2003).
22. Holland, E. C., Hively, W. P., DePinho, R. A. & Varmus, H. E. A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes Dev.* **12**, 3675–3685 (1998).
23. Holland, E. C. Gliomagenesis: genetic alterations and mouse models. *Nature Rev. Genet.* **2**, 120–129 (2001).
24. Bachoo, R. M. *et al.* Epidermal growth factor receptor and Ink4a/Arf: convergent mechanisms governing terminal differentiation and transformation along the neural stem cell to astrocyte axis. *Cancer Cell* **1**, 269–277 (2002).
25. Oliver, T. G. & Wechsler-Reya, R. J. Getting at the root and stem of brain tumors. *Neuron* **42**, 885–888 (2004).
26. Potgens, A. J., Bolte, M., Huppertz, B., Kaufmann, P. & Frank, H. G. Human trophoblast contains an intracellular protein reactive with an antibody against CD133—a novel marker for trophoblast. *Placenta* **22**, 639–645 (2001).
27. Bayani, J. *et al.* Molecular cytogenetic analysis of medulloblastomas and supratentorial primitive neuroectodermal tumors by using conventional banding, comparative genomic hybridization, and spectral karyotyping. *J. Neurosurg.* **93**, 437–448 (2000).
28. Galli, R. *et al.* Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res.* **64**, 7011–7021 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Borden, J. Ma, I. Diplock, M. Ho and C. Gibson for technical assistance, and we are grateful to V. Bonn, L. Davidson, N. Lifshitz and J. Chen of the Mouse Imaging Centre for help with neuroimaging. We thank J. Dick for discussions. S. Singh was supported by a Terry Fox Foundation fellowship from the Canadian Cancer Society, the Neurosurgical Research and Education Foundation and the American Brain Tumor Association. This work was supported by the Canadian Cancer Society, the Canadian Institutes of Health Research, the Foundation of The Hospital for Sick Children, BrainChild and the Jack Baker and Jessica Durigon family funds.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.D. (peter.dirks@sickkids.ca).

A FADD-dependent innate immune mechanism in mammalian cells

Siddharth Balachandran*, Emmanuel Thomas* & Glen N. Barber

Department of Microbiology and Immunology and Sylvester Comprehensive Cancer Center, University of Miami School of Medicine, Miami, Florida 33136, USA

* These authors contributed equally to this work

Vertebrate innate immunity provides a first line of defence against pathogens such as viruses and bacteria. Viral infection activates a potent innate immune response, which can be triggered by double-stranded (ds)RNA produced during viral replication^{1–3}. Here, we report that mammalian cells lacking the death-domain-containing protein FADD^{4,5} are defective in intracellular dsRNA-activated gene expression, including production of type I (α/β) interferons, and are thus very susceptible to viral infection. The signalling pathway incorporating FADD is largely independent of Toll-like receptor 3 and the dsRNA-dependent kinase PKR, but seems to require receptor interacting protein 1 as well as Tank-binding kinase 1-mediated activation of the transcription factor IRF-3. The requirement for FADD in mammalian host defence is evocative of innate immune signalling in

Drosophila, in which a FADD-dependent pathway responds to bacterial infection by activating the transcription of antimicrobial genes⁶. These data therefore suggest the existence of a conserved pathogen recognition pathway in mammalian cells that is essential for the optimal induction of type I interferons and other genes important for host defence.

A major consequence of virus infection is the induction of the type I interferons (IFNs), which are a family of cytokines essential for host defence^{3,7}. While investigating the mechanisms of IFN antiviral activity, we noticed that early passage murine embryonic fibroblasts (MEFs) lacking the apoptosis adaptor molecule FADD appeared to be overtly sensitive to virus infection⁸. To examine this phenotype further, we infected *Fadd*^{+/-} and *Fadd*^{-/-} MEFs with the IFN-sensitive prototypic rhabdovirus vesicular stomatitis virus (VSV), and observed that VSV replication was markedly increased (>100-fold) in the *Fadd*^{-/-} MEFs compared with their wild-type counterparts (Fig. 1a–c). Moreover, whereas pre-treatment with type I (α/β) or type II (γ) IFN for 12 h was seen to exert significant antiviral activity in normal cells, these key antiviral cytokines only delayed the onset of viral replication in *Fadd*^{-/-} MEFs for 24–36 h, whereupon virus replication proceeded unchecked (Fig. 1a–c). The observed susceptibility to infection did not appear to be restricted to VSV, because MEFs lacking FADD were also sensitive to infection by other viruses, including influenza virus and encephalomyocarditis (EMCV) virus (Supplementary Fig. 1). MEFs lacking caspase 8 (refs 9, 10) (or cells treated with caspase inhibitors) exhibited no increased susceptibility to VSV infection compared to control cells, and retained the ability to be protected by IFN, suggesting that FADD exerts its antiviral effects independently of caspase 8 (Supplementary Fig. 2 and data not shown).

Because exposure to type I and II IFNs was unable to effectively protect *Fadd*^{-/-} MEFs from virus replication, it is plausible that effective IFN-mediated, janus kinase (JAK)-activated STAT (signal transducer and activator of transcription) signalling¹¹ may require FADD. However, neither nuclear translocation of STAT1 nor the induction of selected type I and II IFN-responsive genes appeared to be impaired in *Fadd*^{-/-} cells, indicating that IFN signalling *per se* is probably not compromised (Fig. 1d; see also Supplementary Fig. 3).

Despite these observations, it remained possible that the antiviral state initially established by 12 h of pre-exposure to exogenous IFN is short-lived and may require constant *de novo* synthesis of type I IFNs after virus infection (Fig. 1a). Indeed, we noted that continuous supplementation of recombinant type I IFNs to *Fadd*^{-/-} cells after VSV infection conferred some protection against viral replication and cytolysis (Fig. 1e, f). The continual requirement for IFN induction after virus infection was further emphasized by demonstrating that antibody-mediated neutralization of secreted IFN- α or - β after exposure to VSV rendered normal cells susceptible to infection even after IFN pre-treatment (Fig. 1g, h).

These analyses suggest that an actual defect in the production of type I IFNs after virus/dsRNA detection might be responsible for the susceptibility of *Fadd*^{-/-} cells to infection. To examine this possibility, we transfected *Fadd*^{+/-} and *Fadd*^{-/-} cells with a luciferase reporter construct under control of the IFN- β promoter and subsequently administered poly(I:C), a synthetic mimetic of viral dsRNA thought to be the primary trigger of IFN production after virus infection³. Notably, we found that transfected poly(I:C) triggered robust (~10-fold) induction of the IFN- β promoter in *Fadd*^{+/-} cells but not in *Fadd*^{-/-} cells (Fig. 2a). The induction of the IFN- β promoter was not observed using non-transfected, exogenous poly(I:C) alone (Fig. 2a). Poly(I:C)-induced activation of IFN- β promoter-driven luciferase activity could be partially restored by reintroducing murine FADD into *Fadd*^{-/-} MEFs (Supplementary Fig. 4). A requirement for FADD in dsRNA-induced activation of the IFN- β promoter was further confirmed by short interfering RNA (siRNA)-mediated suppression of FADD in HeLa cells (Supplementary Fig. 4).

To investigate the role of FADD in dsRNA-induced gene expression, we subjected dsRNA-transfected or untreated *Fadd*^{+/+} and *Fadd*^{-/-} MEFs to DNA microarray analysis. Over 39,000 transcripts were examined at 0, 3, 12 and 24 h after transfection. At 3 h after transfection, we noticed that the expression of type I IFN genes, especially *Ifna2*, *Ifna4* and *Ifna5*, as well as several other key IFN-inducible genes, including *Irf7*, were markedly reduced in *Fadd*^{-/-} cells compared with controls (Fig. 2b). By 12 h after transfection, the induction of type I IFN-dependent IFN genes had increased in *Fadd*^{-/-} cells, but remained significantly below the levels seen in control cells (data not shown). The impairment of type I IFN induction in the absence of FADD was confirmed at the messenger RNA level by polymerase chain reaction with reverse transcription (RT-PCR), and at the protein level by enzyme-linked immunosorbent assay (ELISA) (Fig. 2c–e). Interestingly, although induction of IFN- β -Luc was severely defective in the absence of FADD, some endogenous *Ifnb* transcriptional activity remained evident, possibly accounting for the eventual, albeit impaired, induction of type I IFN-dependent genes observed by DNA microarray analysis (Fig. 2a–c and data not shown). This may indicate the existence of FADD-independent dsRNA signalling events, which target elements in the endogenous IFN- β promoter independent of those found in the IFN- β -Luc construct (Fig. 2d, e). Nevertheless, given that Fig. 1g, h demonstrates a requirement for *de novo* type I IFN production for efficient antiviral activity, these data show that a defect in the production of type I IFNs in *Fadd*^{-/-} MEFs may explain their susceptibility to virus infection.

Because the induction of type I IFNs was not readily observed using non-transfected, exogenous poly(I:C) alone (Fig. 2a, c–e), we surmised that IFN production in wild-type MEFs may involve intracellular dsRNA recognition molecules such as PKR^{12–15}. However, MEFs lacking PKR largely retained the ability to induce type I IFN and IFN-inducible genes in response to transfected dsRNA (Supplementary Fig. 5). Furthermore, it has been shown that Toll-like receptor 3 (TLR3) is involved in the recognition of extracellular dsRNA¹⁶, an event that can lead to the induction of IFN- β through the adaptor molecule TRIF/TICAM-1 (refs 17, 18). We, however,

did not observe significant type I IFN induction in MEFs or HeLa cells in response to exogenous dsRNA. Indeed, our data show that siRNA-mediated depletion of FADD—but not TLR3 or PKR (or both simultaneously)—in HeLa cells results in the reduction of IFN- β promoter activity in response to transfected poly(I:C) (Supplementary Figs 4 and 5). These results are in agreement with previous findings demonstrating effective type I IFN induction by intracellular dsRNA in PKR- or TLR3-deficient cells^{19–22}, thus suggesting the existence of alternative TLR3/PKR-independent dsRNA signalling pathways in mammalian cells.

To determine whether FADD has a downstream role in TLR signalling, we co-transfected *Fadd*^{+/+} or *Fadd*^{-/-} MEFs with an IFN- β -luciferase reporter construct and plasmids encoding various components of this signalling pathway. However, we observed no abrogation of TLR3-mediated induction of the IFN- β promoter in *Fadd*^{-/-} cells (Fig. 2f). These results were verified by demonstrating that TRAF6- or TRIF/TICAM-1-deficient MEFs, unlike *Fadd*^{-/-} MEFs, retain the ability to resist viral infection after IFN pre-treatment (data not shown). Collectively, this would suggest that FADD is not significantly involved in TLR3-activated, TRIF/TICAM-1- or TRAF6-dependent signalling.

Intriguingly, FADD has recently been reported to be involved in the innate immune response to bacterial infection in *Drosophila*^{23,24}. In this organism, the *imd* gene product, a homologue of the mammalian DD-containing kinase RIP1, associates with *Drosophila* (d)FADD to trigger activation of an NF- κ B-related pathway and the subsequent induction of antibacterial genes⁶. To determine whether an IMD-like pathway involving FADD exists in mammalian cells, we infected IFN-treated or untreated early-passage wild-type (*Ripk1*^{+/+}) and RIP1-deficient (*Ripk1*^{-/-}) MEFs²⁵ with VSV. We found that VSV induced cytolysis in *Ripk1*^{-/-} cells even in the presence of IFN, similar to *Fadd*^{-/-} MEFs (Fig. 3a, b). Approximately 10- to 50-fold more VSV was generated in IFN-treated *Ripk1*^{-/-} MEFs compared with wild-type MEFs (Supplementary Fig. 6). Although IFN-mediated JAK/STAT signalling was found to be intact, *Ripk1*^{-/-} MEFs, as well as HeLa cells in which RIP1 expression was abrogated using RNA interference, exhibited a selective and evident inability to respond to intracellular dsRNA-

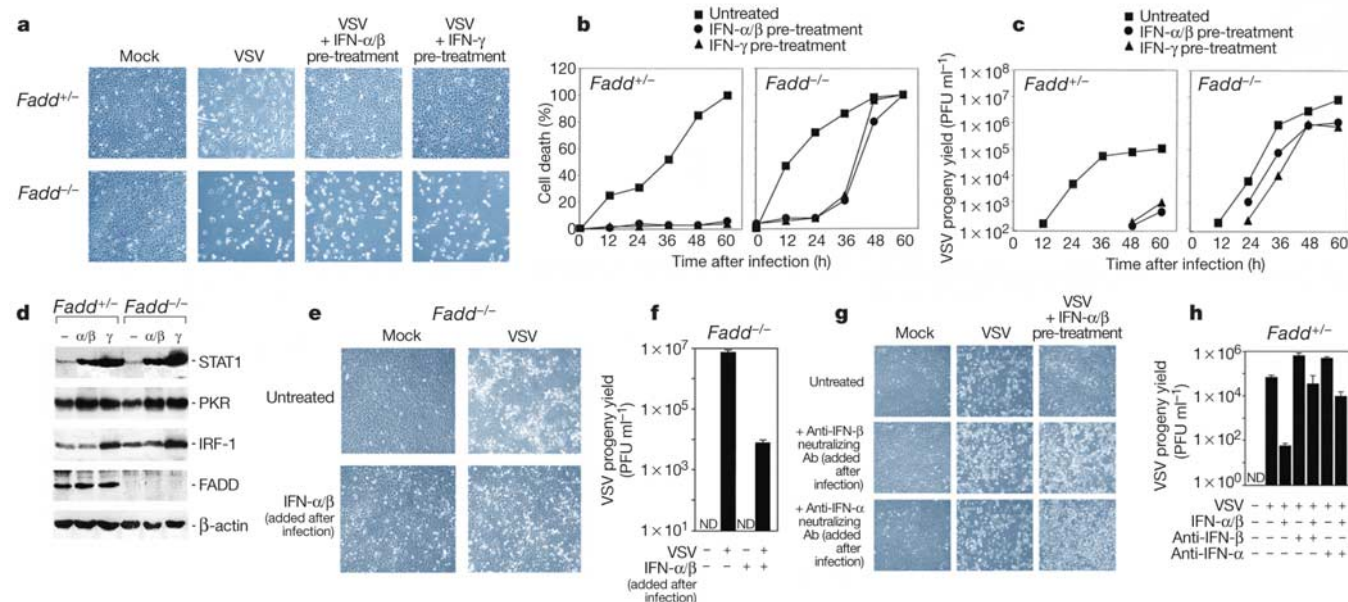


Figure 1 FADD-deficient MEFs are susceptible to VSV despite IFN pre-treatment. **a**, Infection of *Fadd*^{+/+} and *Fadd*^{-/-} MEFs with VSV (multiplicity of infection (MOI) = 10) after IFN- α/β (100 U ml⁻¹) or IFN- γ (0.5 ng ml⁻¹) pre-treatment. **b**, Viability of infected cells. **c**, Progeny virion production from infected cells. PFU, plaque-forming units. **d**, Induction of IFN-responsive genes in IFN-treated *Fadd*^{+/+} and *Fadd*^{-/-} MEFs.

e, Protection of *Fadd*^{-/-} MEFs by continuous supplementation of IFN- α/β (100 U ml⁻¹). **f**, Progeny virion production from infected cells. ND, not detectable. **g**, Neutralizing IFN- α or IFN- β after VSV infection renders *Fadd*^{+/+} cells susceptible to infection despite IFN pre-treatment. **h**, Progeny virion production from infected cells. Error bars are mean $\pm$ s.d.

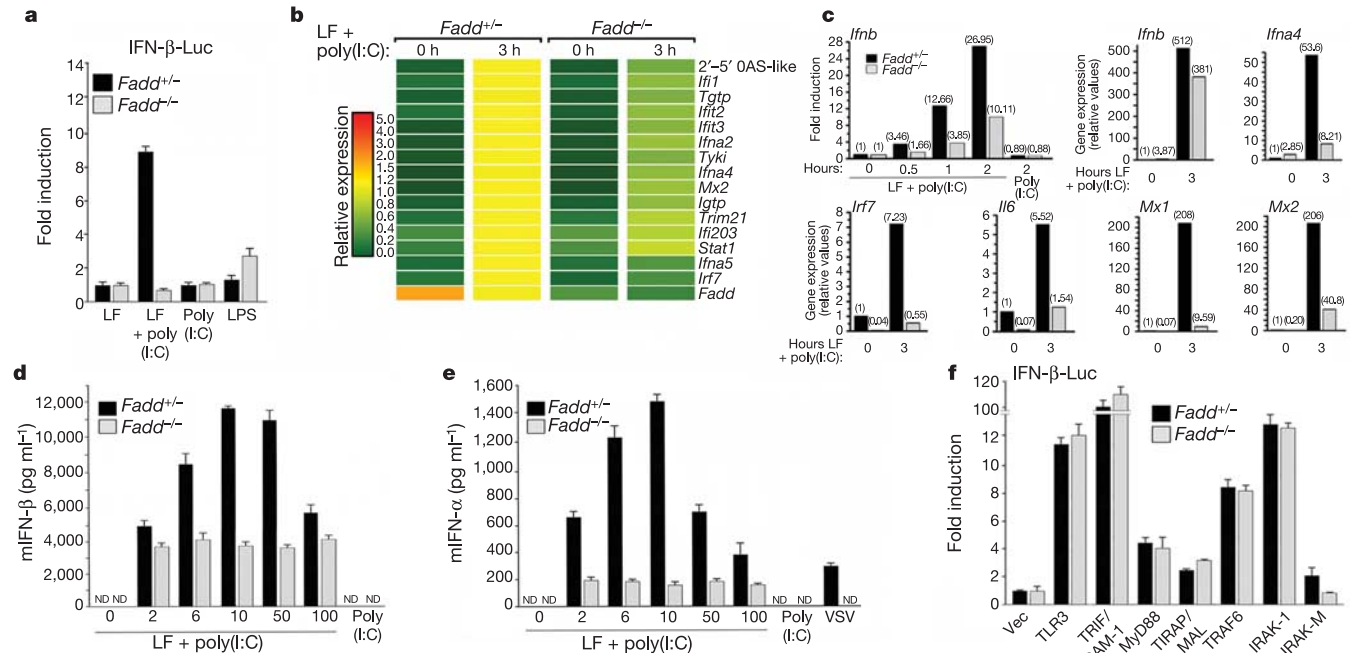


Figure 2 Defective antiviral gene induction by intracellular dsRNA in the absence of FADD. **a**, Defective IFN- β promoter activation by transfected poly(I:C) in *Fadd*^{-/-} MEFs. LF, Lipofectamine2000. **b**, DNA microarray analysis of a selected set of antiviral genes. Induced levels of these genes in *Fadd*^{+/-} cells were arbitrarily set to 1 (yellow) in this and other experiments. **c**, Relative mRNA expression levels of a selection of antiviral genes measured by real time RT-PCR. A detailed time course analysis of IFN- β gene induction in

Fadd^{+/-} versus *Fadd*^{-/-} cells is also shown (top left panel). **d**, IFN- β production after transfection of poly(I:C), or treatment with poly(I:C) alone. **e**, IFN- α production after transfection with poly(I:C), treatment with poly(I:C) alone, or infection with VSV (MOI = 50). **f**, Normal TLR signalling in the absence of FADD. All error bars indicate mean $\pm$ s.d.

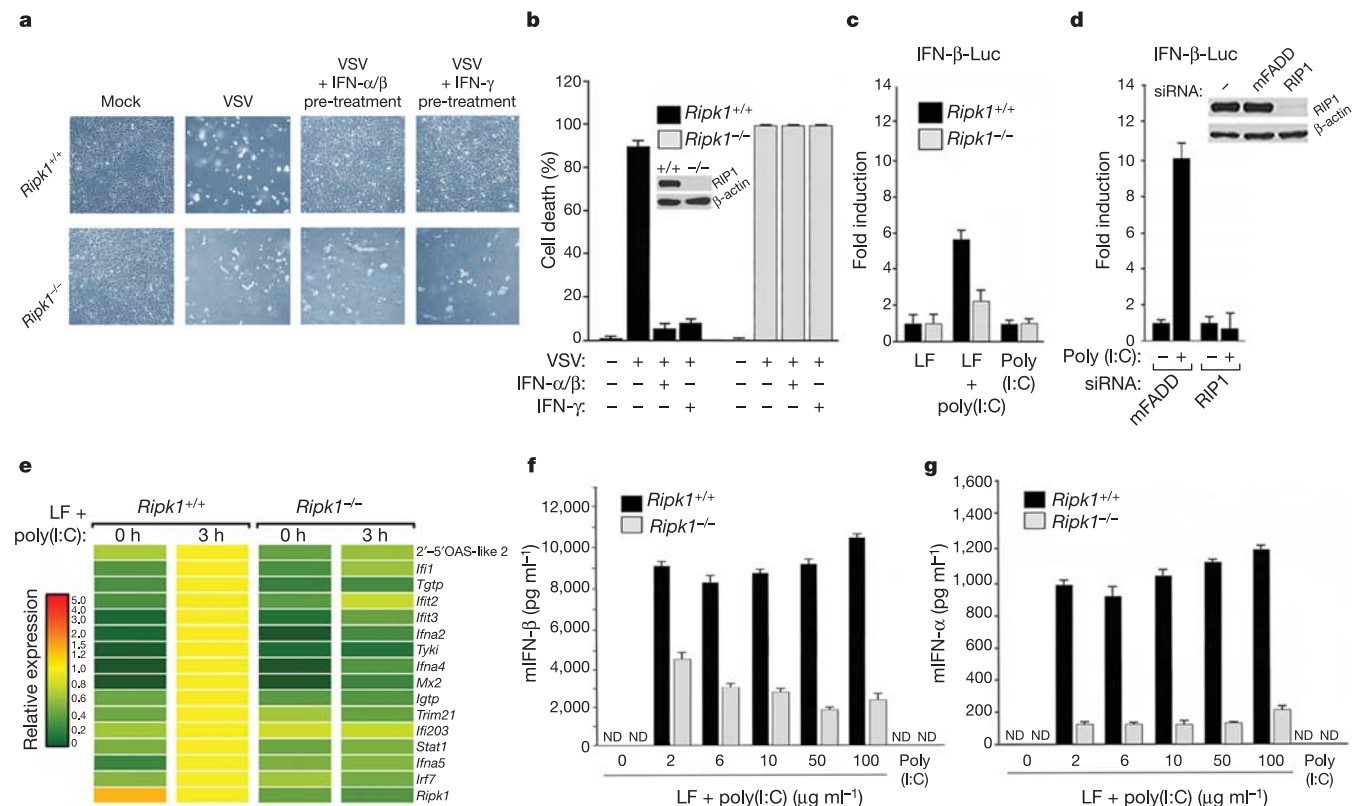


Figure 3 RIP1 deficiency mimics FADD ablation. **a**, Infection of *Ripk1*^{+/+} and *Ripk1*^{-/-} MEFs with VSV (MOI = 10) after pre-treatment with IFN- α/β (100 U ml⁻¹) or IFN- γ (0.5 ng ml⁻¹). **b**, Viability of infected cells. **c**, Defective IFN- β promoter activation by transfected poly(I:C) in *Ripk1*^{-/-} MEFs. **d**, Defective IFN- β promoter activation by

transfected poly(I:C) (2 μ g ml⁻¹) in RIP1 siRNA-treated HeLa cells. **e**, DNA microarray analysis of a selected set of antiviral genes. **f**, IFN- β production after transfection with poly(I:C), or treatment with poly(I:C) alone. **g**, IFN- α production after transfection with poly(I:C), or treatment with poly(I:C) alone. Error bars indicate mean $\pm$ s.d.

mediated activation of the IFN- β promoter (Fig. 3c, d; see also Supplementary Fig. 6). Overexpression of TLR3, IRAK1 and TRAF6 retained the ability to stimulate the IFN- β promoter, suggesting that RIP1 is probably not involved in the TLR3 arm of IFN- β induction, in agreement with other studies²⁶ (Supplementary Fig. 6). DNA microarray analysis also confirmed that *Ripk1*^{-/-} cells exhibit a defect in type I IFN production after exposure to intracellular dsRNA (Fig. 3e). Notably, a lack of type I IFN induction was confirmed at the protein level by IFN- β and IFN- α ELISA (Fig. 3f, g). Thus, RIP1 seems to be required for efficient intracellular, dsRNA-mediated induction of type I IFNs.

In *Drosophila*, IMD and dFADD are required to stimulate the induction of antimicrobial gene expression through activation of the NF- κ B homologue Relish via an I- κ B kinase (IKK) complex comprised of IKK- β /IRD5 and IKK- γ /Kenny⁶. In mammalian cells, induction of IFN- β also involves IKK-mediated activation of NF- κ B, as well as IRF-3 (ref. 27). However, pre-treatment with IFN was able to effectively protect MEFs lacking IKK- α , - β or - γ against virus infection (Fig. 4a). This study was complemented by examining MEFs lacking Tank-binding kinase 1 (TBK-1; also known as

IKK- δ , NAK and T2K), as this molecule seems to be the primary IRF-3 kinase in MEFs^{19,28,29}. This experiment revealed that, similar to *Fadd*^{-/-} and *Ripk1*^{-/-} fibroblasts, TBK-1-deficient cells are not protected against virus replication and cytolysis even after pre-treatment with IFN (Fig. 4a; see also Supplementary Fig. 7). Similar to the situation in *Fadd*^{-/-} cells, these data could be explained by a defect in type I IFN induction in TBK-1-deficient MEFs. Indeed, DNA microarray, RT-PCR and ELISA analyses confirmed a severe impairment of dsRNA-responsive induction of type I IFN, as well as other antiviral genes, in the absence of TBK-1, in agreement with other studies (Fig. 4b–d and data not shown)^{19,30}. These data indicate that FADD may mediate its effects predominantly through TBK-1 activation of IRF-3. Accordingly, IRF-3 translocation, which occurs after phosphorylation by TBK-1 and IKK- ϵ ^{28,29}, was found to be defective in *Fadd*^{-/-} cells after treatment with transfected dsRNA (Fig. 4e, f). Furthermore, *Irf3*^{-/-} MEFs were not fully protected against virus infection after exposure to type I or II IFNs (Fig. 4g; see also Supplementary Fig. 7). Similarly, DNA microarray, RT-PCR, ELISA and RNA interference analyses confirmed a defect in the ability of intracellular dsRNA to induce type I IFN production in

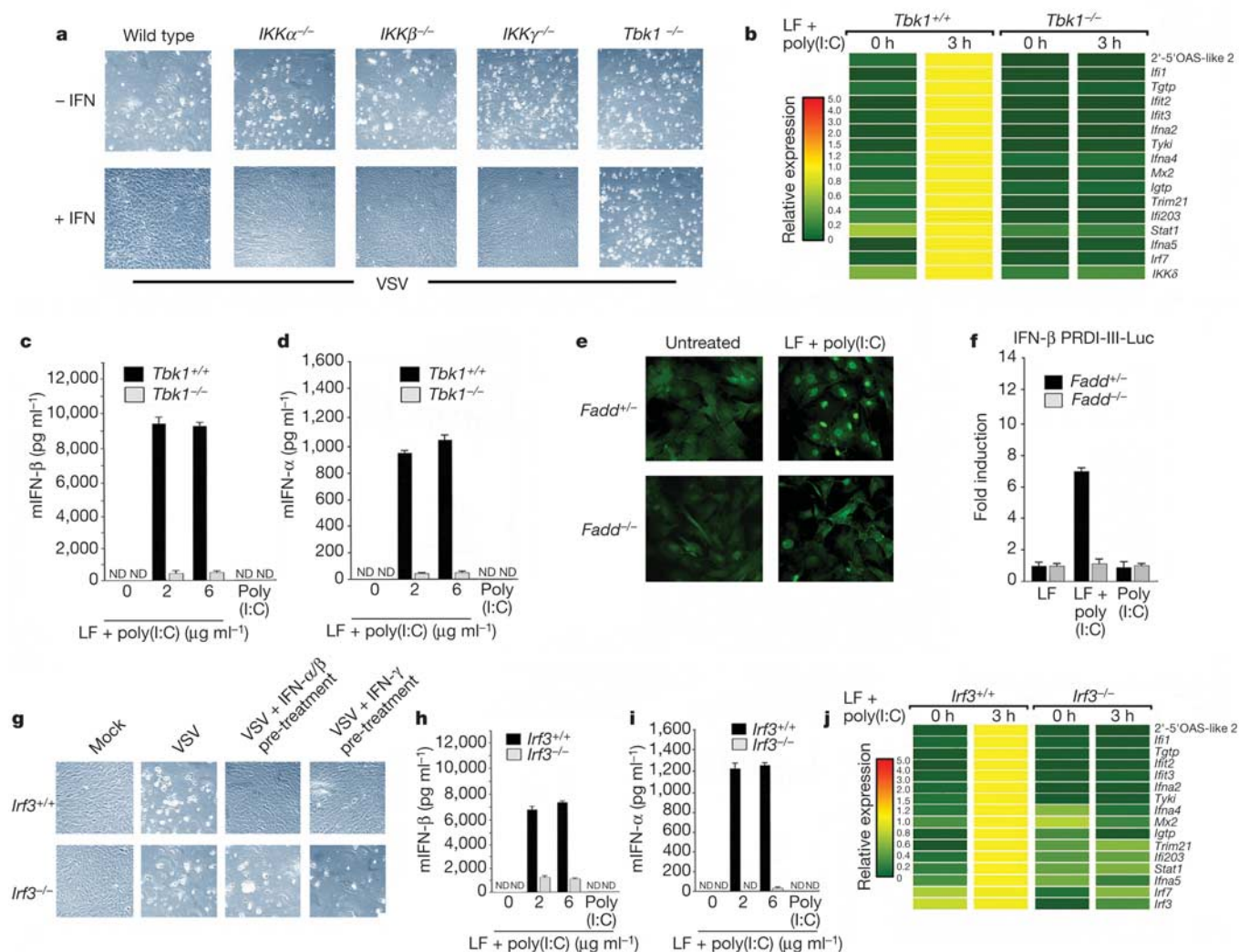


Figure 4 The antiviral pathway incorporating FADD signals via TBK-1 and IRF-3. **a**, Infection of wild-type or IKK- α -, IKK- β -, IKK- γ - and TBK-1-deficient MEFs with VSV (MOI = 10) with or without IFN- α/β (100 U ml⁻¹) pre-treatment. **b**, DNA microarray analysis of a selected set of antiviral genes. **c**, IFN- β production after transfection with poly(I:C), or treatment with poly(I:C) alone. **d**, IFN- α production after transfection with the indicated amounts of poly(I:C), or treatment with poly(I:C) alone. **e**, Localization of IRF-3 after transfection of poly(I:C) for 1 h in *Fadd*^{+/+} and *Fadd*^{-/-} cells. **f**, Defective IRF-3-

responsive promoter activation in *Fadd*^{-/-} MEFs. **g**, Infection of *Irf3*^{+/+} and *Irf3*^{-/-} MEFs with VSV (MOI = 10) with or without IFN- α/β (100 U ml⁻¹) or IFN- γ (0.5 ng ml⁻¹) pre-treatment. **h**, IFN- β production after transfection with poly(I:C), or treatment with poly(I:C) alone. **i**, IFN- α production after transfection with poly(I:C), or treatment with poly(I:C) alone. **j**, DNA microarray analysis for a selected set of antiviral genes. Error bars indicate mean $\pm$ s.d.

Irf3^{-/-} MEFs (Fig. 4h–k, Supplementary Fig. 8 and data not shown).

These results suggest that viral dsRNAs are recognized by an intracellular receptor molecule, which may recruit FADD and RIP1 into an ‘innateosome’ complex to regulate TBK-1-mediated activation of IRF-3. We show that the loss of FADD or RIP1 leads to a defect in IFN- β production (and perhaps the induction of other primary immune response genes) and a consequent lag in the production of IRF-7 and members of the IFN- α family, which are necessary for fortification of the antiviral state³. It is also noteworthy that TBK-1-deficient MEFs display a more profound defect in the induction of type I IFNs in response to dsRNA stimulation than either FADD-deficient or RIP1-deficient MEFs alone, plausibly suggesting that intracellular dsRNA-activated complexes retain some activity in the absence of FADD, or that alternative FADD-independent intracellular signalling cascades converge on TBK-1. This RIP1/FADD/TBK-1 (RIFT) pathway seems to be largely independent of TLR3, PKR, TRIF/TICAM-1 or TRAF6, and is in agreement with other findings suggesting the existence of alternative intracellular, dsRNA-activated signal transducers, such as the DExD/H helicase RIG-I (ref. 22) (Supplementary Fig. 9). The requirement for FADD in mammalian innate immunity is reminiscent of the IMD pathway in *Drosophila*, which is important for defence against Gram-negative bacteria⁶. Thus, similar to the Toll pathway, an alternative (FADD-dependent) innate immune pathway may be evolutionarily conserved between insects and mammals. □

Methods

Cells, viruses and reagents

MEFs were provided as follows: *Fadd*^{+/-}, *Fadd*^{-/-}, *Tbk1*^{+/-} and *Tbk1*^{-/-} (W.-C. Yeh); *Casp8*^{+/-} and *Casp8*^{-/-} (D. Wallach); *Traf6*^{+/-} and *Traf6*^{-/-} (J. Inoue); *Ripk1*^{+/-} and *Ripk1*^{-/-} (N. Kelliher); *Pkr*^{+/-} and *Pkr*^{-/-} (J. Bell and B. Williams); *Stat1*^{+/-} and *Stat1*^{-/-} (J. Durbin); *Trif/TICAM-1*^{+/-} and *Trif/TICAM-1*^{-/-} (S. Akira); *IKK*^{+/-}, *IKK α* ^{-/-}, *IKK β* ^{-/-} and *IKK γ* ^{-/-} (M. Karin); and *Irf3*^{+/-} and *Irf3*^{-/-} (K. Mossman and B. Williams). All other cell lines were obtained from the American Type Culture Collection (ATCC). Poly(I:C) (Amersham-Pharmacia) was reconstituted in PBS at 2 mg ml⁻¹, denatured at 55 °C for 30 min, and allowed to anneal to room temperature before use. Unless indicated otherwise, MEFs were transfected with 6 μ g of poly(I:C) in 8 μ l of Lipofectamine2000 or 100 μ g poly(I:C) alone to stimulate IFN- β promoter activation and IFN production in MEFs. Luciferase assays were performed 6 h after treatment. Supernatants for ELISA were collected 24 h after treatment. VSV (Indiana strain), EMCV and influenza virus (strain A/WSN/33) were used in infections and titred by standard techniques. Human IFN- α and IFN- β and murine IFN- α ELISA Kits were acquired from PBL. Murine IFN- β ELISA assays were performed at PBL. z-VAD was from ICN Pharmaceuticals. All other reagents were from Sigma, unless mentioned otherwise.

Plasmids

Plasmids were obtained from the following sources: GFP-STAT1 (N. Reich); murine (m)FADD (ATCC); TICAM-1/TRIF, IRAK1, IRAK-M, TLR3, MyD88, TIRAP/MAL and TRAF6 (all from Invivogen); PRD I-III-Luc (T. Maniatis); IFN- β -Luc, ISRE-Luc and GAS-Luc (J. Hiscott).

RNA interference

Chemically synthesized 21-nucleotide sense and antisense RNA oligonucleotides were obtained from Dharmacon. HeLa and 293 cells were plated on six-well plates at 60,000 and 200,000 cells per well, respectively, and transfected with 100 pmol of siRNA duplex per well using Oligofectamine (Invitrogen). Assays were typically performed 72 h after siRNA treatment, when gene knockdown was found to be maximal. TLR3 and PKR siRNAs were obtained as SMARTpool proprietary sequences. Other siRNAs were as follows: hFADD, GAAGACCUGUGUGCAGCAU; mFADD, ACGAUCUGAUGGAGCUCAA; RIP1, GGCGAAGAUGAUGAACAGAU; IRF-3, AAUCUACGAGUUUGUGAAC.

Antibodies

Antibodies were obtained from the following sources: FADD and pSTAT1 (Upstate); PKR and IRF-1 (Santa Cruz); RIP1, STAT1 and Fas (Pharmingen); β -actin (Sigma); TRAF6 (Stressgen); TBK-1/IKK- δ (Imgenex); IRF-3 (Zymed). Neutralizing antibodies to mIFN- β and mIFN- α were purchased from Research Diagnostics.

DNA microarray analysis

Total RNA was extracted from MEFs stimulated with or without poly(I:C) (6 μ g ml⁻¹ in Lipofectamine2000). Preparation of cDNA and microarray analysis was performed at the W.M. Keck Foundation Biotechnology Research Laboratory DNA microarray facility at Yale University. The mouse genome 430 2.0 array (Affymetrix) was used. Data analysis was performed with Microarray Suite software (version 5.0; Affymetrix) and GeneSpring software (Silicon Genetics).

Real-time PCR

Total RNA was isolated from cells using the RNeasy RNA extraction kit (Qiagen) and cDNA synthesis was performed using 1 μ g of total RNA (Roche). Fluorescence real-time PCR analysis was performed using a LightCycler 2.0 instrument (Roche Molecular Biochemicals) and TaqMan gene expression assays (Applied Biosystems). Relative amounts of mRNA were normalized to the 18S ribosomal RNA levels in each sample.

Received 1 August; accepted 15 October 2004; doi:10.1038/nature03124.

1. Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu. Rev. Immunol.* **21**, 335–376 (2003).
2. Medzhitov, R. & Janeway, C. Jr. The Toll receptor family and microbial recognition. *Trends Microbiol.* **8**, 452–456 (2000).
3. Taniguchi, T. & Takaoka, A. The interferon- α/β system in antiviral responses: a multimodal machinery of gene regulation by the IRF family of transcription factors. *Curr. Opin. Immunol.* **14**, 111–116 (2002).
4. Chinnaiyan, A. M., O'Rourke, K., Tewari, M. & Dixit, V. M. FADD, a novel death domain-containing protein, interacts with the death domain of Fas and initiates apoptosis. *Cell* **81**, 505–512 (1995).
5. Yeh, W. C. *et al.* FADD: essential for embryo development and signaling from some, but not all, inducers of apoptosis. *Science* **279**, 1954–1958 (1998).
6. Hoffmann, J. A. The immune response of *Drosophila*. *Nature* **426**, 33–38 (2003).
7. Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H. & Schreiber, R. D. How cells respond to interferons. *Annu. Rev. Biochem.* **67**, 227–264 (1998).
8. Balachandran, S. *et al.* Alpha/beta interferons potentiate virus-induced apoptosis through activation of the FADD/Caspase-8 death signaling pathway. *J. Virol.* **74**, 1513–1523 (2000).
9. Varfolomeev, E. E. *et al.* Targeted disruption of the mouse Caspase 8 gene ablates cell death induction by the TNF receptors, Fas/Apo1, and DR3 and is lethal prenatally. *Immunity* **9**, 267–276 (1998).
10. Muzio, M. *et al.* FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death-inducing signaling complex. *Cell* **85**, 817–827 (1996).
11. Levy, D. E. & Darnell, J. E. Jr. Stats: transcriptional control and biological impact. *Nature Rev. Mol. Cell Biol.* **3**, 651–662 (2002).
12. Williams, B. R. PKR: a sentinel kinase for cellular stress. *Oncogene* **18**, 6112–6120 (1999).
13. Balachandran, S. *et al.* Essential role for the dsRNA-dependent protein kinase PKR in innate immunity to viral infection. *Immunity* **13**, 129–141 (2000).
14. Chu, W. M. *et al.* JNK2 and IKK β are required for activating the innate response to viral infection. *Immunity* **11**, 721–731 (1999).
15. Diebold, S. S. *et al.* Viral infection switches non-plasmacytoid dendritic cells into high interferon producers. *Nature* **424**, 324–328 (2003).
16. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF- κ B by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
17. Oshiumi, H., Matsumoto, M., Funami, K., Akazawa, T. & Seya, T. TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon- β induction. *Nature Immunol.* **4**, 161–167 (2003).
18. Yamamoto, M. *et al.* Role of adaptor TRIF in the MyD88-independent toll-like receptor signaling pathway. *Science* **301**, 640–643 (2003).
19. Hemmi, H. *et al.* The roles of two I κ B kinase-related kinases in lipopolysaccharide and double stranded RNA signaling and viral infection. *J. Exp. Med.* **199**, 1641–1650 (2004).
20. Honda, K. *et al.* Selective contribution of IFN- α/β signaling to the maturation of dendritic cells induced by double-stranded RNA or viral infection. *Proc. Natl Acad. Sci. USA* **100**, 10872–10877 (2003).
21. Hoebe, K. *et al.* Upregulation of costimulatory molecules induced by lipopolysaccharide and double-stranded RNA occurs by Trif-dependent and Trif-independent pathways. *Nature Immunol.* **4**, 1223–1229 (2003).
22. Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nature Immunol.* **5**, 730–737 (2004).
23. Leulier, F., Vidal, S., Saigo, K., Ueda, R. & Lemaitre, B. Inducible expression of double-stranded RNA reveals a role for dFADD in the regulation of the antibacterial response in *Drosophila* adults. *Curr. Biol.* **12**, 996–1000 (2002).
24. Naitza, S. *et al.* The *Drosophila* immune defense against gram-negative infection requires the death protein dFADD. *Immunity* **17**, 575–581 (2002).
25. Kelliher, M. A. *et al.* The death domain kinase RIP mediates the TNF-induced NF- κ B signal. *Immunity* **8**, 297–303 (1998).
26. Meylan, E. *et al.* RIP1 is an essential mediator of Toll-like receptor 3-induced NF- κ B activation. *Nature Immunol.* **5**, 503–507 (2004).
27. Wathlet, M. G. *et al.* Virus infection induces the assembly of coordinately activated transcription factors on the IFN- β enhancer *in vivo*. *Mol. Cell* **1**, 507–518 (1998).
28. Sharma, S. *et al.* Triggering the interferon antiviral response through an IKK-related pathway. *Science* **300**, 1148–1151 (2003).
29. Fitzgerald, K. A. *et al.* IKK ϵ and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunol.* **4**, 491–496 (2003).
30. McWhirter, S. M. *et al.* IFN-regulatory factor 3-dependent gene expression is defective in *Tbk1*-deficient mouse embryonic fibroblasts. *Proc. Natl Acad. Sci. USA* **101**, 233–238 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to W.-C. Yeh, N. Kelliher, D. Wallach, J. Inoue, J. Durbin, J. Bell, K. Mossman, E. Harhaj, M. Karin, B. Williams and S. Akira for fibroblasts, and T. Maniatis, J. Hiscott and N. Reich and for plasmid constructs. We also thank G. Spruill, M. Fallahi and T. Andrew for technical assistance. This work was supported by DARPA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.N.B. (gbarber@med.miami.edu).

Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks

Yentram Huyen^{1,2}, Omar Zgheib^{1,2}, Richard A. DiTullio Jr.^{1,2}, Vassilis G. Gorgoulis^{1,4}, Panayotis Zacharatos^{1,4}, Tom J. Petty^{1,2}, Emily A. Sheston¹, Hestia S. Mellert¹, Elena S. Stavridi¹ & Thanos D. Halazonetis^{1,3}

¹Wistar Institute, Philadelphia, Pennsylvania 19104-4268, USA

²Biomedical Graduate Studies Program and ³Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

⁴Department of Histology and Embryology, School of Medicine, University of Athens, Athens, 11527, Greece

The mechanisms by which eukaryotic cells sense DNA double-strand breaks (DSBs) in order to initiate checkpoint responses are poorly understood. 53BP1 is a conserved checkpoint protein with properties of a DNA DSB sensor^{1–5}. Here, we solved the structure of the domain of 53BP1 that recruits it to sites of DSBs. This domain consists of two tandem tudor folds with a deep pocket at their interface formed by residues conserved in the budding yeast Rad9 and fission yeast Rhp9/Crb2 orthologues. *In vitro*, the 53BP1 tandem tudor domain bound histone H3 methylated on Lys 79 using residues that form the walls of the

pocket; these residues were also required for recruitment of 53BP1 to DSBs. Suppression of DOT1L, the enzyme that methylates Lys 79 of histone H3, also inhibited recruitment of 53BP1 to DSBs. Because methylation of histone H3 Lys 79 was unaltered in response to DNA damage, we propose that 53BP1 senses DSBs indirectly through changes in higher-order chromatin structure that expose the 53BP1 binding site.

The region of human 53BP1 responsible for its recruitment to sites of DNA DSBs maps to residues 1480–1616 (refs 6, 7). This region, which is conserved in putative 53BP1 orthologues in *Saccharomyces cerevisiae* (Rad9), *Schizosaccharomyces pombe* (Rhp9/Crb2) and *Caenorhabditis elegans* (Hsr-9/TO5F1) (Fig. 1a; refs 8–11), contains a folded domain (residues 1486–1602 of human 53BP1) whose three-dimensional structure we determined by X-ray crystallography at a resolution of 2.8 Å. The structure, which was also recently solved by NMR spectroscopy¹², revealed that the domain consists of ten β-strands and a carboxy-terminal α-helix (Fig. 1b). The amino-terminal five β-strands and the C-terminal five β-strands adopt folds that are identical to each other and to the fold of the tudor domain of the survival motor neuron (SMN) protein^{13,14}. SMN has only one tudor fold, but in 53BP1 two tandem tudor folds comprise a single globular domain.

The residues of 53BP1 that are conserved in the 53BP1/Rad9 family were mapped on the three-dimensional structure. Some of the conserved residues (coloured blue in Fig. 1a) are important for folding. The remaining conserved residues (coloured magenta) map predominantly to one surface of the domain and mostly to a deep pocket located at the interface of the two tudor folds (Figs 1b and 2a).

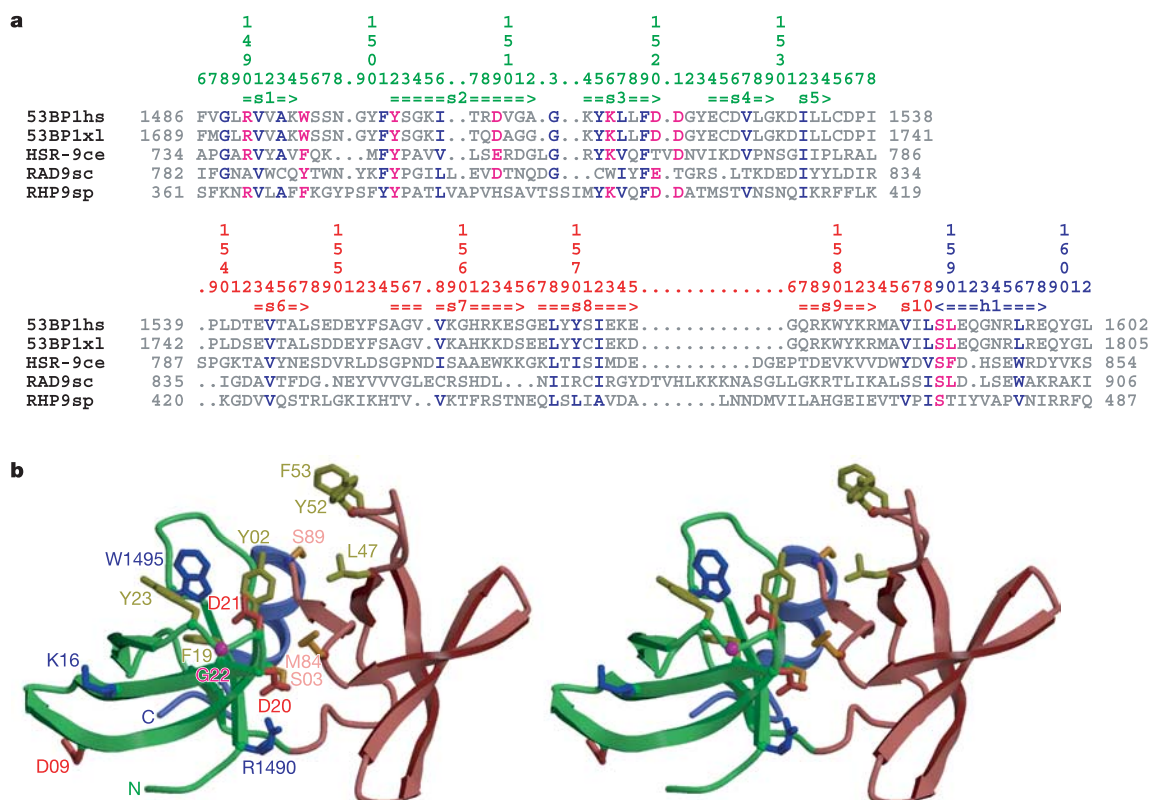


Figure 1 Evolutionary conservation and three-dimensional structure of the human 53BP1 tandem tudor domain. **a**, Sequence conservation of residues 1486–1602 of human 53BP1. Conserved residues are coloured blue or magenta, depending on whether they appear to contribute to folding (blue) or not (magenta). Codon numbers along the top refer to the human sequence and are coloured green, red and blue for N-terminal tudor fold, C-terminal tudor fold and C-terminal helix, respectively. Secondary structure elements (s, strand; h, helix) are also indicated. 53BP1hs, human 53BP1; 53BP1xl, *Xenopus laevis*

53BP1; HSR-9ce, *C. elegans* Hsr-9; RAD9sc, *S. cerevisiae* Rad9; RHP9sp, *S. pombe* Rhp9/Crb2. The sequence alignment was guided by the three-dimensional structure of human 53BP1. **b**, Stereo ribbons representation of the three-dimensional structure of residues 1486–1602. The N-terminal fold is coloured red, the C-terminal tudor fold green and the C-terminal α-helix blue. Residues are labelled using the single-letter amino acid code and the codon number (except for residues 1500–1599, where only the last two digits are shown).

The four walls of this pocket are formed by Trp 1495, Tyr 1502, Met 1584 and Leu 1547, and Asp 1521, respectively, whereas the bottom is formed by Ser 1503 (Fig. 2a). Thus, the residues lining the walls of this deep pocket are hydrophobic with the exception of Asp 1521.

To identify functionally important elements of the 53BP1 tandem tudor domain we used the structure to design amino acid substitutions that would change the surface properties of the domain without compromising folding. Trp 1495 was substituted with Val; Tyr 1502 with Gln or Leu; and Asp1521 with Arg or Ala. In addition, Tyr 1552 and Phe 1553 were substituted with Ala. The last two residues are not evolutionarily conserved, but their aromatic side chains are exposed to solvent (Fig. 1b), reminiscent of aromatic residues in single-stranded DNA-binding proteins that intercalate between DNA bases¹⁵. All the substitutions described above were introduced into a fusion protein containing green fluorescent protein (GFP) and residues 1220–1711 of human 53BP1. The ability of these proteins to localize to sites of DSBs was monitored in live cells 15 min after exposure to 3 Gy ionizing radiation. Wild-type 53BP1 fusion protein and the Tyr1552Ala and Phe1553Ala mutants were recruited efficiently to sites of DSBs. In contrast, all of the substitutions that targeted the pocket at the interface of the tudor folds compromised or abolished recruitment of 53BP1 to DSBs (Fig. 2b). We further examined the effect of the Asp1521Arg substitution in the context of a GFP–53BP1 fusion that contained just the tudor domain and nuclear localization signal of 53BP1 (residues 1480–1711), and in the context of GFP fused to full-length 53BP1 (residues 1–1972). In both cases the Asp1521Arg substitution abolished recruitment to DSBs (Fig. 2b). The effect of the various amino acid substitutions was not due to unfolding of the tudor domain, as ascertained by gel filtration analysis of purified wild-type and mutant polypeptides expressed in *Escherichia coli* (data not shown). We therefore conclude that the deep pocket at the interface of the two tudor folds is the critical structural element for targeting 53BP1 to sites of DNA DSBs.

The tudor domain of SMN interacts with methylated arginine residues present in spliceosomal Sm proteins^{13,14,16,17}. We therefore reasoned that the deep pocket of 53BP1 might interact with

methylated arginine or lysine residues, in which case the physiologically relevant binding partners of 53BP1 might be methylated histones¹⁸. To explore this hypothesis we tested whether 53BP1 would bind calf thymus histones. Histones H2A, H2B, H3 and H4 were all present in the input fraction, but 53BP1 bound histone H3 predominantly (Fig. 3a). The interaction was observed under stringent conditions (1 M KCl and 0.5% Triton X-100) and involved amounts of histone H3 that were readily detected by Coomassie staining. A second protein that migrated slightly slower than histone H4 also bound human 53BP1 (Fig. 3a). N-terminal amino acid sequencing and immunoblotting with various histone H3-specific antibodies revealed that this protein was a cleaved form of histone H3 corresponding to residues 28–135 (Fig. 3b; data not shown). Cleaved histone H3 was also present in the input fraction (Fig. 3b).

Unlike calf thymus histone H3, bacterially expressed histone H3 failed to interact with 53BP1, suggesting that binding required post-translational modification of histone H3 (Fig. 3a). To identify the relevant modification(s), tryptic peptides of full-length and cleaved histone H3 bound to 53BP1 were compared by tandem mass spectrometry to tryptic peptides of full-length histone H3 from the input fraction. The only post-translational modification identified in the 53BP1-bound cleaved histone H3 was methylation of Lys 79. Peptides with non-methylated Lys 79 were not detected in this sample. Lys 79 was also exclusively methylated in the full-length histone H3 bound to 53BP1, whereas non-methylated Lys 79 was readily detectable in histone H3 from the input fraction (Fig. 3c). These results suggest that 53BP1 recognizes histone H3 methylated on Lys 79. Indeed, a synthetic peptide corresponding to residues 74–83 of human histone H3 with di-methylated Lys 79 competed with histone H3 for binding to 53BP1. The corresponding non-methylated peptide did not compete, whereas the mono-methylated Lys 79 peptide and peptides with di-methylated Lys 27 or Arg 26 competed with lower efficiency (Fig. 3d). Binding of the 53BP1 tandem tudor domain to a histone H3 peptide with di-methylated Lys 79 was also demonstrated by isothermal titration calorimetry, which showed a dissociation constant below 1 μ M (data not shown).

If binding to histone H3 was important for recruitment of 53BP1

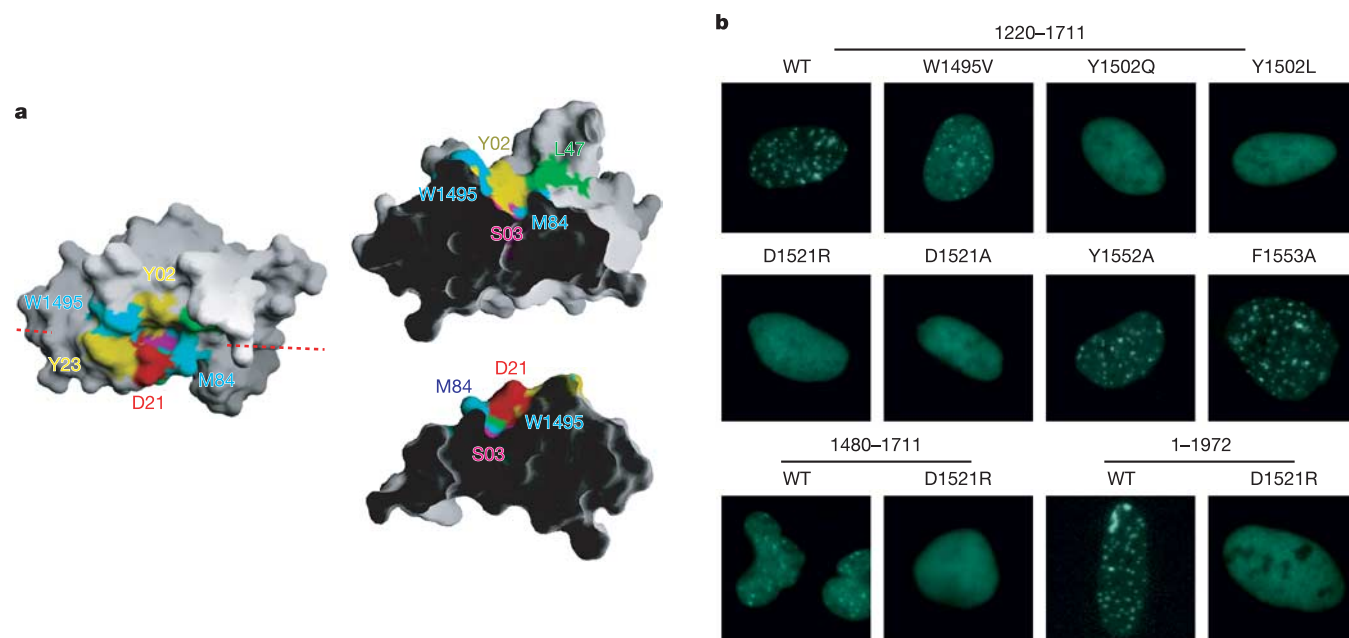


Figure 2 Mapping functionally important 53BP1 residues on the surface of its tandem tudor domain. **a**, Left panel: surface representation of the tandem tudor domain shows the pocket formed by conserved residues. Right panels: surface representations of the domain sliced along the dotted red line in the left panel reveals the depth of the pocket.

The orientation of the domain in the upper panel is that shown in Fig. 1b. The residues, whose projections are coloured on the 53BP1 surface, are labelled as in Fig. 1b. **b**, Intracellular localization of GFP fused to residues 1220–1711, 1480–1711 or 1–1972 of wild-type (WT) or mutant 53BP1 in irradiated cells.

to sites of DNA DSBs, then the interaction should be evolutionarily conserved. Furthermore, 53BP1 mutants that fail to localize to sites of DSBs should not interact with histone H3. Indeed, the *S. cerevisiae* Rad9 DNA damage checkpoint protein, which has the least sequence similarity to human 53BP1 among all of the 53BP1/Rad9 family members (Fig. 1a), interacted with histone H3 (Fig. 3e). Furthermore, the amino acid substitutions that inhibited recruitment of 53BP1 to sites of DNA DSBs (Fig. 2b) also inhibited binding to histone H3 (Fig. 3f).

The interaction between 53BP1 and histone H3 could also be demonstrated *in vivo* by crosslinking. Non-transfected U2OS osteosarcoma cells and transfected U2OS cells expressing His-tagged GFP-53BP1 fusion proteins with wild-type or mutant (Asp1521Arg) 53BP1 sequences were pre-extracted with Triton-X100 *in situ* 15 min after irradiation and then treated with formaldehyde to crosslink interacting proteins. His-tagged GFP-53BP1 was then affinity-purified on nickel-coated beads and bound histone H3 was detected by immunoblotting after the crosslinks were reversed by boiling. Histone H3 was captured on beads incubated with extracts from His-tagged GFP-53BP1(WT) (wild-type 53BP1)-expressing cells treated with formaldehyde, but not on

beads incubated with extracts from cells that were mock-crosslinked or that expressed the Asp1521Arg mutant protein (Fig. 3g). Furthermore, the crosslinking was specific for histone H3 because histone H2A, although present in the Triton-X100 insoluble material, was not captured by the beads (data not shown).

If 53BP1 is recruited to sites of DNA DSBs by interacting with histone H3 methylated on Lys79, then in the absence of Lys79 methylation 53BP1 should fail to form ionizing radiation-induced foci. The enzyme that methylates Lys 79 in human cells is DOT1L, an evolutionarily conserved methyltransferase^{19–21}. Using short interfering RNA (siRNA) specific for human *DOT1L* we suppressed methylation of histone H3 on Lys 79 and also recruitment of 53BP1 to sites of DSBs (Fig. 4a). We did not examine whether *DOT1L* suppression led to checkpoint defects, because the formation of 53BP1 foci was not suppressed in all cells (presumably due to incomplete suppression of Lys 79 methylation). However, deletion of *DOT1* in *S. cerevisiae* leads to radiation sensitivity and a DNA DSB checkpoint defect^{22,23}. At least the radiation sensitivity is due to loss of histone H3 Lys 79 methylation, because substitution of Lys 79 with Ala, Gln or Pro has the same phenotype as deletion of *DOT1* (ref. 22).

The interaction between 53BP1 and histone H3 methylated on

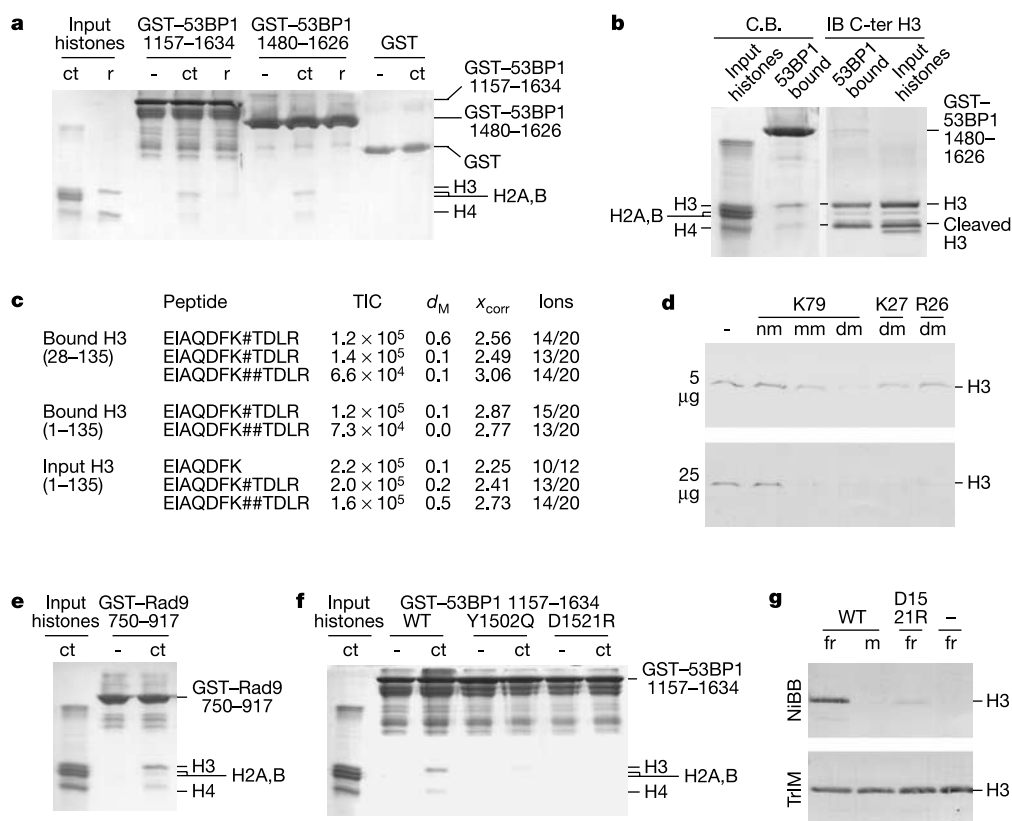


Figure 3 Binding of 53BP1 tandem tudor domain to histone H3 methylated on Lys 79. **a**, GST fusion proteins containing residues 1157–1634 or 1480–1626 of human 53BP1 or GST protein alone were examined for binding to calf thymus (ct) histones H2A, H2B, H3 and H4 or to recombinant (r) histones H3 and H4. **b**, 53BP1-bound histones were identified as full-length or N-terminally-cleaved histone H3 by Coomassie brilliant blue staining (C.B.) or immunoblotting (IB) with an antibody that recognizes the C terminus (C-ter) of histone H3. **c**, 53BP1-bound full-length histone H3 (1–135), 53BP1-bound cleaved histone H3 (28–135) and full-length histone H3 from the input fraction were analysed by tandem mass spectrometry (MS/MS). For each sample, the identified peptides that include Lys 79 are shown. K, K# and K## represent non-, mono- and di-methylated lysine, respectively. TIC, total ion current (an indicator of peptide abundance); d_M , absolute difference of the experimental and theoretical masses; x_{corr} , cross correlation value of the experimental MS/MS spectrum versus the theoretical; Ions, number of matched ions in the experimental MS/MS spectrum versus the total number of theoretically possible ions. **d**, Binding of a GST fusion protein containing 53BP1 residues

1480–1626 to histones prepared from 293T cells was performed in the presence of no competitor peptide (–) or 5 or 25 μ g competitor histone H3 peptides. Bound histone H3 was detected by immunoblotting. Peptides spanning residues 74–83 of histone H3 had either non- (nm), mono- (mm), or di-methylated (dm) Lys 79 (K79); peptides spanning residues 23–34 and 23–32 of histone H3 had di-methylated Lys 27 (K27) and Arg 26 (R26), respectively. **e**, Binding of a GST fusion protein containing *S. cerevisiae* Rad9 residues 750–917 to histone H3 was examined as in **a**. **f**, Amino acid substitutions that disrupt recruitment of 53BP1 to sites of DNA DSBs also disrupt binding to histone H3. Binding was examined as in **a**. **g**, Crosslinking of ectopically expressed His-tagged GFP-53BP1 to endogenous histone H3 in irradiated U2OS cells. Non-transfected (–) U2OS cells or cells transfected with His-tagged GFP-53BP1 fusion proteins (either wild-type or D1521R mutant) were irradiated then crosslinked with formaldehyde (fr) or were mock-crosslinked (m). Triton-X100 insoluble material (TrIM) was incubated with nickel-coated beads. Histone H3 in the TrIM and nickel bead-bound (NiBB) fractions was detected by immunoblotting after reversal of crosslinking.

Lys79 suggests two possibilities with regard to how 53BP1 is recruited to sites of DSBs: either Lys79 becomes methylated at sites of DNA DSBs, or Lys79 is constitutively methylated and its accessibility to 53BP1 changes in response to DNA damage. To distinguish between these two possibilities we prepared histones from non-irradiated and irradiated 293T human carcinoma cells. Irradiation did not lead to enhanced methylation of Lys79 of histone H3, even though phosphorylation of histone H2AX, a known marker of irradiation²⁴, was enhanced (Fig. 4b). Furthermore, histone H3 prepared from irradiated and non-irradiated cells bound equally well to 53BP1 *in vitro* (Fig. 4b). These results suggest that increased exposure of pre-existing methylated Lys79, rather

than newly methylated Lys79, accounts for recruitment of 53BP1 to sites of DNA DSBs.

Increased exposure to pre-existing methylated Lys79 could result from changes in higher-order chromatin structure. Indeed, it has been proposed that DSBs induce long-range changes in chromatin structure, perhaps as a result of relaxation of DNA supercoiling induced by the break²⁵. One of the agents that affects chromatin structure and activates the DNA DSB checkpoint kinase ATM without inducing DNA breaks is mild hypotonic media²⁵. To explore whether changes in chromatin structure are sufficient to target 53BP1 to chromatin, U2OS osteosarcoma cells expressing GFP fused to residues 1220–1711 of 53BP1 were incubated in

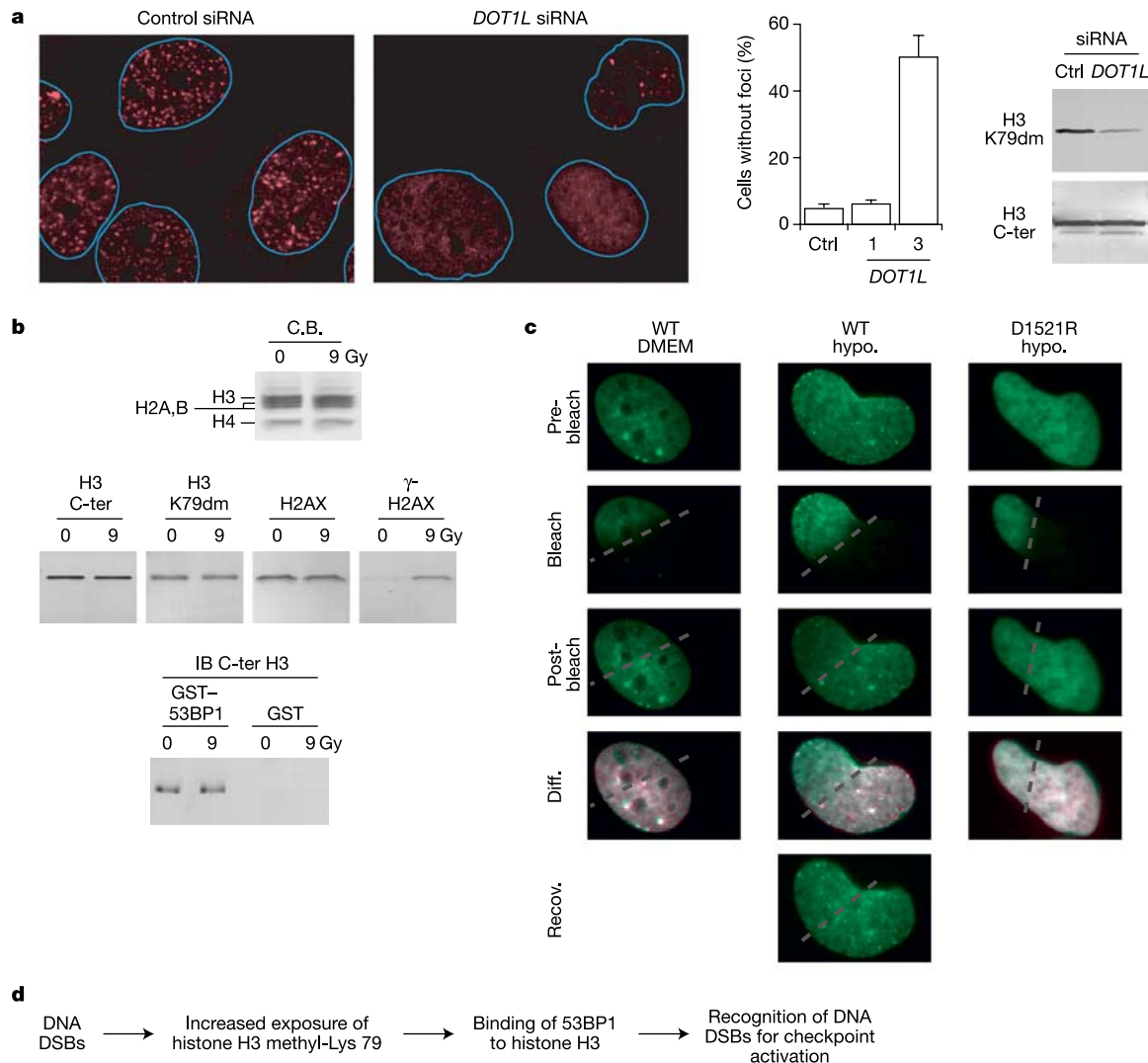


Figure 4 Methylation of histone H3 on Lys79 and changes in higher-order chromatin structure recruit 53BP1 to sites of DSBs and chromatin. **a**, Suppression of histone H3 methylation on Lys79 inhibits recruitment of 53BP1 to sites of DSBs. The data were determined from four independent experiments of control (Ctrl) siRNA-treated cells or cells transfected once (1) or three consecutive times (3) with *DOT1L*-specific siRNA. Bars indicate standard errors. The efficiency of suppression of histone H3 Lys79 methylation after the third round of *DOT1L* siRNA transfection was monitored by immunoblotting with antibodies that recognize the C terminus (C-ter) of histone H3 or histone H3 di-methylated on Lys79 (K79dm). **b**, Irradiation does not enhance methylation of Lys79 of histone H3 or binding of extracted histone H3 to GST–53BP1. Histones prepared from non-irradiated or irradiated (9 Gy) 293T cells were stained with Coomassie brilliant blue (C.B.; top panel) or immunoblotted with antibodies that recognize the C terminus of histone H3, histone H3 di-methylated on Lys79 (K79dm), histone H2AX and phosphorylated histone H2AX (γ -H2AX; middle panel). The histones from the non-irradiated and irradiated cells were also

examined for binding to a GST fusion protein containing human 53BP1 residues 1480–1626 and to plain GST as a control (bottom panel). Bound histone H3 was detected by immunoblotting. **c**, Exposure of cells to hypotonic media slows the nuclear diffusion of 53BP1. Cells expressing GFP fused to residues 1220–1711 of 53BP1 that were cultured in regular (DMEM) or hypotonic (hypo.) media were photographed (pre-bleach image), then a portion of a cell's nucleus (bleach image) was bleached for 10 min and another image (post-bleach) was acquired at the end of the bleach period. To show differences in GFP–53BP1 fluorescence before and after bleaching, the post-bleach image was pseudocoloured magenta and merged with the pre-bleach image (Diff.). The GFP–53BP1(WT)-expressing cells exposed to hypotonic media were allowed to recover from the photobleaching for 7 min and then another image was acquired (Recov.). Note that the cells cultured in hypotonic media have altered nuclear architecture, as indicated by the morphology of the nucleoli. **d**, Model for recognition of DNA DSBs by 53BP1.

regular tissue culture media or for 1 h in hypotonic media, and nuclear diffusion of GFP–53BP1 was monitored by photobleaching. We expected that changes in higher-order chromatin structure would target GFP–53BP1 to chromatin and slow its diffusion through the nucleus. A portion of the nucleus of a live cell was bleached with light for 10 min and GFP–53BP1 fluorescence was examined immediately thereafter. In the control cells GFP–53BP1 fluorescence was equal in the bleached and non-bleached areas, indicating fast kinetics of 53BP1 diffusion during bleaching, whereas in the cells exposed to hypotonic media GFP–53BP1 fluorescence was lower in the bleached area, indicating slower 53BP1 diffusion (Fig. 4c). The decrease in diffusion kinetics was dependent on wild-type 53BP1, because the GFP–53BP1 (Asp1521Arg) mutant protein diffused with fast kinetics in cells exposed to hypotonic media (Fig. 4c). We conclude that changes in chromatin structure are sufficient to recruit 53BP1 to chromatin.

Very little is known regarding how cells sense the presence of DNA DSBs in order to activate DNA damage checkpoint pathways. It had been proposed that the tandem tudor domain of 53BP1 binds DNA directly⁷. However, the highly purified protein that we used for crystallization did not bind DNA (data not shown). It had also been suggested that 53BP1 is recruited to sites of DSBs by binding to phosphorylated histone H2AX (γ -H2AX)⁶. Indeed, a region of 53BP1 that is N-terminal to the tandem tudor domain binds γ -H2AX *in vitro*⁶, and retention of 53BP1 at sites of DSBs requires histone H2AX²⁶. However, the initial recruitment of 53BP1 to sites of DSBs is not defective in *H2AX*^{-/-} cells²⁶, and the tandem tudor domain, which is sufficient for targeting 53BP1 to sites of DSBs, does not bind γ -H2AX *in vitro*⁶.

Our studies suggest that 53BP1 targeting to sites of DNA DSBs is mediated by interaction of its tandem tudor domain with methylated Lys 79 of histone H3. Lys 79 of histone H3 is constitutively methylated in both mammalian and yeast cells^{19–21} and at least in human cells irradiation did not lead to increased histone H3 Lys 79 methylation. Therefore, DNA DSBs probably affect the accessibility of methylated Lys 79 of histone H3 to 53BP1, possibly through changes in higher-order chromatin structure²⁵. Histone H3 Lys 79 maps to the histone core²⁷ (Supplementary Fig. 1) and would be inaccessible to 53BP1 if higher-order chromatin structure involves nucleosome stacking, as recently proposed²⁸. Disruption of nucleosome stacking by DSBs would lead to exposure of methylated Lys 79 of histone H3, as well as of any other methylated residues in the histone core (such as histone H4 Lys20, T. Kouzarides, personal communication), resulting in recognition of the DSB by 53BP1 (Fig. 4d). Further analysis is needed to establish whether the mechanism is as simple as proposed here or whether we have just scratched the surface regarding how cells sense the presence of DNA DSBs for activation of the checkpoint. □

Methods

Structure determination

A polypeptide corresponding to residues 1483–1624 of human 53BP1 was expressed in *E. coli* and purified by anion exchange (Sephacrose Q and then Resource Q columns; Pharmacia) and gel filtration (Superdex 200 column; Pharmacia) chromatography. The protein (50 mg ml⁻¹) in buffer containing 25 mM BTP, pH 6.8, 200 mM KCl and 5 mM dithiothreitol (DTT) was mixed with an equal volume of reservoir solution containing 18–21% PEG 3350 and 200 mM magnesium nitrate, pH 5.8, and crystals containing ten molecules per unit cell were grown at 4 °C by the hanging-drop vapour diffusion method. Heavy atom derivatives were obtained by soaking the crystals in reservoir solution supplemented with 1.25 mM thimerosal and 5–25% glycerol. Data collection and determination of the structure (Supplementary Table 1) were performed as previously described²⁹.

Intracellular localization of GFP–53BP1 fusion proteins

U2OS osteosarcoma cells transiently expressing GFP fused to residues 1220–1711, 1480–1711 or 1–1972 of wild-type or mutant human 53BP1 were exposed to 3 Gy ionizing radiation and examined 15 min later by live fluorescence microscopy. To study diffusion kinetics, cells expressing GFP fused to residues 1220–1711 of 53BP1 were cultured in regular media (DMEM supplemented with 10% fetal calf serum) or were switched 1 h before bleaching to hypotonic media (PBS with 100 mM NaCl supplemented with 0.45%

glucose (w/v) and 1% serum). Part of a cell was bleached for 5 or 10 min using the mercury lamp of the microscope (100 W, Leica) and by partially closing its field diaphragm. During the recovery phase the cell was not illuminated. While on the microscope the cells were maintained at 37 °C using stage and objective lens heaters (Biopetechs). The results shown are representative of three independent experiments with at least three cells examined per experiment.

Histone H3 binding

Glutathione S-transferase (GST) proteins (5 μ g) fused to residues 1157–1634 or 1480–1626 of human 53BP1, or 750–917 of *S. cerevisiae* Rad9 were bound to glutathione beads (Pharmacia). The beads were then incubated for 1 h at 4 °C with histones in buffer containing 25 mM BTP, pH 6.8, 1 M KCl and 0.5% Triton X-100. After six washes the bound material was resolved on SDS-polyacrylamide gels and either stained with Coomassie brilliant blue or immunoblotted with an antibody that recognizes the C terminus of histone H3 (AbCam). The histones for these studies were derived from calf thymus (50 μ g, Worthington), 293T cells (40 μ g) or were expressed in *E. coli* (5 μ g each of histones H3 and H4, Upstate). Histones from 293T cells were prepared by lysis in buffer containing 50 mM Tris-HCl, pH 8.0, 120 mM NaCl, 0.5% NP40, 1 mM DTT and protease-kinase-phosphatase inhibitors^{1,5} for 45 min at 4 °C. After centrifugation the pellet was incubated in buffer containing 10 mM HEPES, 1.5 mM MgCl₂, 10 mM KCl, 0.5 mM DTT, 1.5 mM PMSF and 0.25 N HCl for 1 h at 4 °C and the extracted histones were neutralized by adding one-fifth volume 1.5 M Tris-HCl, pH 8.8. For tandem mass spectrometry analysis, protein tryptic peptides were resolved on a reverse-phase liquid chromatography column, which was directly coupled to a quadrupole ion trap mass spectrometer, and the data were interpreted with SEQUEST software. Binding of GST–53BP1 to histones in the presence of competitor peptides (AbCam) was performed in buffer containing 25 mM BTP, pH 6.8, 250 mM KCl and 0.5% Triton X-100 for 1 h at 4 °C.

Crosslinking

Proteins were crosslinked with formaldehyde³⁰. Parental U2OS cells and U2OS cells expressing N-terminally His-tagged GFP–53BP1 fusion proteins containing residues 1220–1711 of 53BP1 were exposed to 9 Gy ionizing radiation. Fifteen minutes later the cells were washed with PBS and incubated on ice for 20 min with Triton-X100 extraction buffer (10 mM PIPES, pH 6.8, 100 mM NaCl, 300 mM sucrose, 3 mM MgCl₂, 1 mM EGTA, 0.2% Triton X-100 and protease-kinase-phosphatase inhibitors). The cells were then rinsed with PBS, incubated for 10 min at 4 °C with 1% formaldehyde in PBS or just with PBS (mock-crosslinked) and then for 5 min with 0.1 M glycine. After a PBS wash, the cells were collected by scraping and centrifuged. The pellet was incubated for 15 min on ice with lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% SDS and inhibitors), disrupted by sonication and diluted tenfold in buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.2% Triton X-100 and inhibitors. The disrupted pellet, hereafter referred to as Triton-X100 insoluble material, was incubated with Ni-NTA magnetic agarose beads (Qiagen) and bound proteins were eluted with imidazole. Histone H3 was detected by immunoblotting after the crosslinks were reversed by boiling.

Suppression of histone H3 methylation on Lys 79

U2OS cells were transfected with control siRNA (luciferase; Dharmacon) or siRNA directed against human *DOT1L* (5'-GCUCGCUAUGGAGAAUACdTdT-3'), as previously described³, except that for some cells the siRNA transfections were repeated every 4 days for a total of three rounds of transfections, because of the long half-life of histones and because histone methylation may not be reversible. Three days after the last siRNA transfection, the cells were exposed to 9 Gy ionizing radiation and 53BP1 intracellular localization was monitored by immunofluorescence 15 min later³.

Received 6 July; accepted 15 October 2004; doi:10.1038/nature03114.

Published online 3 November 2004.

- Schultz, L. B., Chehab, N. H., Malikzay, A. & Halazonetis, T. D. p53 binding protein 1 (53BP1) is an early participant in the cellular response to DNA double-strand breaks. *J. Cell Biol.* **151**, 1381–1390 (2000).
- Xia, Z., Morales, J. C., Dunphy, W. G. & Carpenter, P. B. Negative cell cycle regulation and DNA damage-inducible phosphorylation of the BRCT protein 53BP1. *J. Biol. Chem.* **276**, 2708–2718 (2001).
- Anderson, L., Henderson, C. & Adachi, Y. Phosphorylation and rapid relocalization of 53BP1 to nuclear foci upon DNA damage. *Mol. Cell Biol.* **21**, 1719–1729 (2001).
- Rappold, I., Iwabuchi, K., Date, T. & Chen, J. Tumor suppressor p53 binding protein 1 (53BP1) is involved in DNA damage-signaling pathways. *J. Cell Biol.* **153**, 613–620 (2001).
- Mochan, T. A., Venere, M., DiTullio, R. A. Jr & Halazonetis, T. D. 53BP1 and NFB1/MDC1-Nbs1 function in parallel interacting pathways activating ataxia-telangiectasia mutated (ATM) in response to DNA damage. *Cancer Res.* **63**, 8586–8591 (2003).
- Ward, I. M., Minn, K., Jorda, K. G. & Chen, J. Accumulation of checkpoint protein 53BP1 at DNA breaks involves its binding to phosphorylated histone H2AX. *J. Biol. Chem.* **278**, 19579–19582 (2003).
- Iwabuchi, K. et al. Potential role for 53BP1 in DNA end-joining repair through direct interaction with DNA. *J. Biol. Chem.* **278**, 36487–36495 (2003).
- Weinert, T. A. & Hartwell, L. H. The *RAD9* gene controls the cell cycle response to DNA damage in *Saccharomyces cerevisiae*. *Science* **241**, 317–322 (1988).
- Willson, J., Wilson, S., Warr, N. & Watts, F. Z. Isolation and characterization of the *Schizosaccharomyces pombe* *rhp9* gene: a gene required for the DNA damage checkpoint but not the replication checkpoint. *Nucleic Acids Res.* **25**, 2138–2146 (1997).
- Saka, Y., Esashi, F., Matsusaka, T., Mochida, S. & Yanagida, M. Damage and replication checkpoint control in fission yeast is ensured by interactions of Crb2, a protein with BRCT motif, with Cut5 and Chk1. *Genes Dev.* **11**, 3387–3400 (1997).
- Boulton, S. J. et al. Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127–131 (2002).

12. Charier, G. *et al.* The tudor tandem of 53BP1: a new structural motif involved in DNA and RG-rich peptide binding. *Structure* **12**, 1551–1562 (2004).
13. Selenko, P. *et al.* SMN tudor domain structure and its interaction with the Sm proteins. *Nature Struct. Biol.* **8**, 27–31 (2001).
14. Sprangers, R., Groves, M. R., Sinning, I. & Sattler, M. High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J. Mol. Biol.* **327**, 507–520 (2003).
15. Theobald, D. L., Mitton-Fry, R. M. & Wuttke, D. S. Nucleic acid recognition by OB-fold proteins. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 115–133 (2003).
16. Friesen, W. J., Massenet, S., Pauskhin, S., Wyce, A. & Dreyfuss, G. SMN, the product of the spinal muscular atrophy gene, binds preferentially to dimethylarginine-containing protein targets. *Mol. Cell* **7**, 1111–1117 (2001).
17. Brahm, H., Meheus, L., de Brabandere, V., Fischer, U. & Luhrmann, R. Symmetrical dimethylation of arginine residues in spliceosomal Sm protein B/B' and the Sm-like protein LSm4, and their interaction with the SMN protein. *RNA* **7**, 1531–1542 (2001).
18. Kouzarides, T. Histone methylation in transcriptional control. *Curr. Opin. Genet. Dev.* **12**, 198–209 (2002).
19. Feng, Q. *et al.* Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr. Biol.* **12**, 1052–1058 (2002).
20. van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**, 745–756 (2002).
21. Lacoste, N., Utley, R. T., Hunter, J. M., Poirier, G. G. & Cote, J. Disruptor of telomeric silencing-1 is a chromatin-specific histone H3 methyltransferase. *J. Biol. Chem.* **277**, 30421–30424 (2002).
22. Game, J. C., Williamson, M. S. & Baccari, C. X-ray survival characteristics and genetic analysis for nine *Saccharomyces* deletion mutants that affect radiation sensitivity. *Genetics* online publication, 15 September 2004 (doi:10.1534/genetics.104.028613).
23. San-Segundo, P. A. & Roeder, G. S. Role for the silencing protein Dot1 in meiotic checkpoint control. *Mol. Biol. Cell* **11**, 3601–3615 (2000).
24. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks *in vivo*. *J. Cell Biol.* **146**, 905–916 (1999).
25. Bakkenist, C. J. & Kastan, M. B. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* **421**, 499–506 (2003).
26. Celeste, A. *et al.* Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5**, 675–679 (2003).
27. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
28. Mozziconacci, J. & Victor, J. M. Nucleosome gaping supports a functional structure for the 30 nm chromatin fiber. *J. Struct. Biol.* **143**, 72–76 (2003).
29. Hyen, Y. *et al.* Structural differences in the DNA binding domains of human p53 and its *C. elegans* ortholog Cep-1. *Structure* **12**, 1237–1243 (2004).
30. Kannouche, P. L., Wing, J. & Lehmann, A. R. Interaction of human DNA polymerase ϵ with monoubiquitinated PCNA: a possible mechanism for the polymerase switch in response to DNA damage. *Mol. Cell* **14**, 491–500 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors thank N. Pavletich, S. Berger, G. Dreyfuss and R. Kaufman for support and discussions, the Wistar Institute Proteomics Facility (K. Speicher) for protein N-terminal sequencing and mass spectrometry analysis, and the University of Pennsylvania Protein Chemistry Facility (J. Lambris and M. Katragadda) for calorimetry analysis. This work was supported by a grant to T.D.H. from the National Cancer Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.D.H. (halazonetis@wistar.upenn.edu). Structure coordinates have been deposited in the Protein Data Bank under the accession code 1XNI.

Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine

Robert T. Batey, Sunny D. Gilbert & Rebecca K. Montange

Department of Chemistry and Biochemistry, 215 UCB, University of Colorado, Boulder, Colorado 80309, USA

Riboswitches are genetic regulatory elements found in the 5' untranslated region of messenger RNA that act in the absence of protein cofactors^{1,2}. They are broadly distributed across bacteria and account for the regulation of more than 2% of all genes in *Bacillus subtilis*, underscoring their importance in the control of

cellular metabolism³. The 5' untranslated region of many mRNAs of genes involved in purine metabolism and transport contain a guanine-responsive riboswitch that directly binds guanine, hypoxanthine or xanthine to terminate transcription^{3,4}. Here we report the crystal structure at 1.95 Å resolution of the purine-binding domain of the guanine riboswitch from the *xpt-pbuX* operon of *B. subtilis* bound to hypoxanthine, a prevalent metabolite in the bacterial purine salvage pathway. This structure reveals a complex RNA fold involving several phylogenetically conserved nucleotides that create a binding pocket that almost completely envelops the ligand. Hypoxanthine functions to stabilize this structure and to promote the formation of a downstream transcriptional terminator element, thereby providing a mechanism for directly repressing gene expression in response to an increase in intracellular concentrations of metabolite.

The remarkable ability of artificially selected RNA 'aptamers' to bind practically any imaginable ligand^{5,6} has been harnessed to engineer RNA-based biosensors and molecular machines that respond to environmental cues⁷. These approaches hold considerable promise for real-life applications⁸. Yet, nature has preceded these efforts, as illustrated by naturally occurring RNA sensors, called 'riboswitches', that directly control gene expression through their ability to bind various small-molecule metabolites^{1,2}. These sensors are exemplified by the guanine-responsive riboswitch that controls the transcription of genes associated with purine metabolism in numerous bacterial species³. The predicted secondary structure of this motif consists of three helices (P1–P3) that surround a three-way junction (Fig. 1a), with phylogenetically conserved nucleotides located in the junction and loops. Immediately downstream of the guanine-binding domain in the mRNA is the switching domain (Fig. 1b), which has been proposed to control gene expression by forming either a terminator or an antiterminator element, depending on whether metabolite is bound^{3,4}.

To understand how this natural biosensor functions, we have solved by X-ray crystallography the structure of the guanine-binding domain bound to hypoxanthine (Supplementary Table S1 and Supplementary Fig. S1), a biologically relevant ligand of the guanine-responsive riboswitch. In the hypoxanthine-bound state, the RNA adopts a three-dimensional fold in which the terminal loops (L2 and L3) form a series of interconnecting hydrogen bonds (see pairing scheme in Fig. 1a) to bring the P2 and P3 helices parallel to each other (Fig. 1c, d). Unfavourable electrostatic interactions, a result of the juxtaposition of regions of the ribose-phosphate backbone, are neutralized through the binding of several cations between the two backbones (Supplementary Fig. S2). Anchoring the global helical arrangement of the RNA are numerous tertiary contacts around the three-way junction, dominated by the J2/3 loop (Fig. 1c) interacting with bound hypoxanthine, the P1 helix, and the J1/2 and J3/1 loops.

The purine-binding pocket is created by conserved nucleotides in and around the three-way junction element. These nucleotides help to define the purine-binding pocket through the formation of two sets of base triples above and below (Fig. 1a). The 3' side of the pocket is flanked by a water-mediated U22–A52–A73 base triple and an A23–G46–C53 triple; in both cases, the Watson–Crick face of the adenosine interacts with the minor groove of a Watson–Crick pair (Fig. 2a). The other side is created by sequential base triples between conserved Watson–Crick pairs at the top of helix P1 (U20–A76 and A21–U75) and the Watson–Crick faces of U49 and C50, respectively, which fasten the J2/3 loop to the P1 helix. This extensive use of base triples to create a ligand-binding site is very similar to *in vitro* selected RNA aptamers that recognize planar ring systems, as exemplified by the structures of the theophylline⁹, FMN¹⁰ and malachite green¹¹ binders. Thus, artificially selected RNAs use some of the same principles for creating binding sites for small-molecule ligands as their naturally occurring counterpart.

Hypoxanthine is bound through an extensive series of hydrogen bonds with nucleotides U22, U47, U51 and C74 (Fig. 2b), forming a base quadruple that stacks directly on the P1 helix. The structure clearly shows that the mRNA contacts all of the functional groups in the ligand, thereby explaining the specificity for hypoxanthine observed in biochemical studies³. In addition, guanine binding can be readily rationalized, because there is room in the structure to accommodate an exocyclic amine at the 2-position of the bound purine. This additional functional group can form hydrogen bonds with the carbonyl oxygens at the 2-position of C74 and U51, consistent with the tenfold higher affinity of this riboswitch for guanine over hypoxanthine³.

One of the most marked features is how the ligand is almost

completely enveloped by the RNA (Fig. 2c): 97.8% of the surface of hypoxanthine is inaccessible to bulk solvent in the complex. The almost complete use of a ligand for recognition by an RNA is unprecedented among structurally characterized aptamers, although selection strategies that do not involve immobilization of the ligand on a solid support¹² hold promise for the development of RNAs that are capable of a similar degree of ligand burial. This finding also implies that the local binding site must undergo a substantial conformational change upon ligand binding, because it is not possible for hypoxanthine to gain access to a preformed binding site, which is a common feature of many RNA–ligand interactions^{13,14}. Finally, this mode of purine recognition also easily explains its ability to change specificity from guanine to adenine

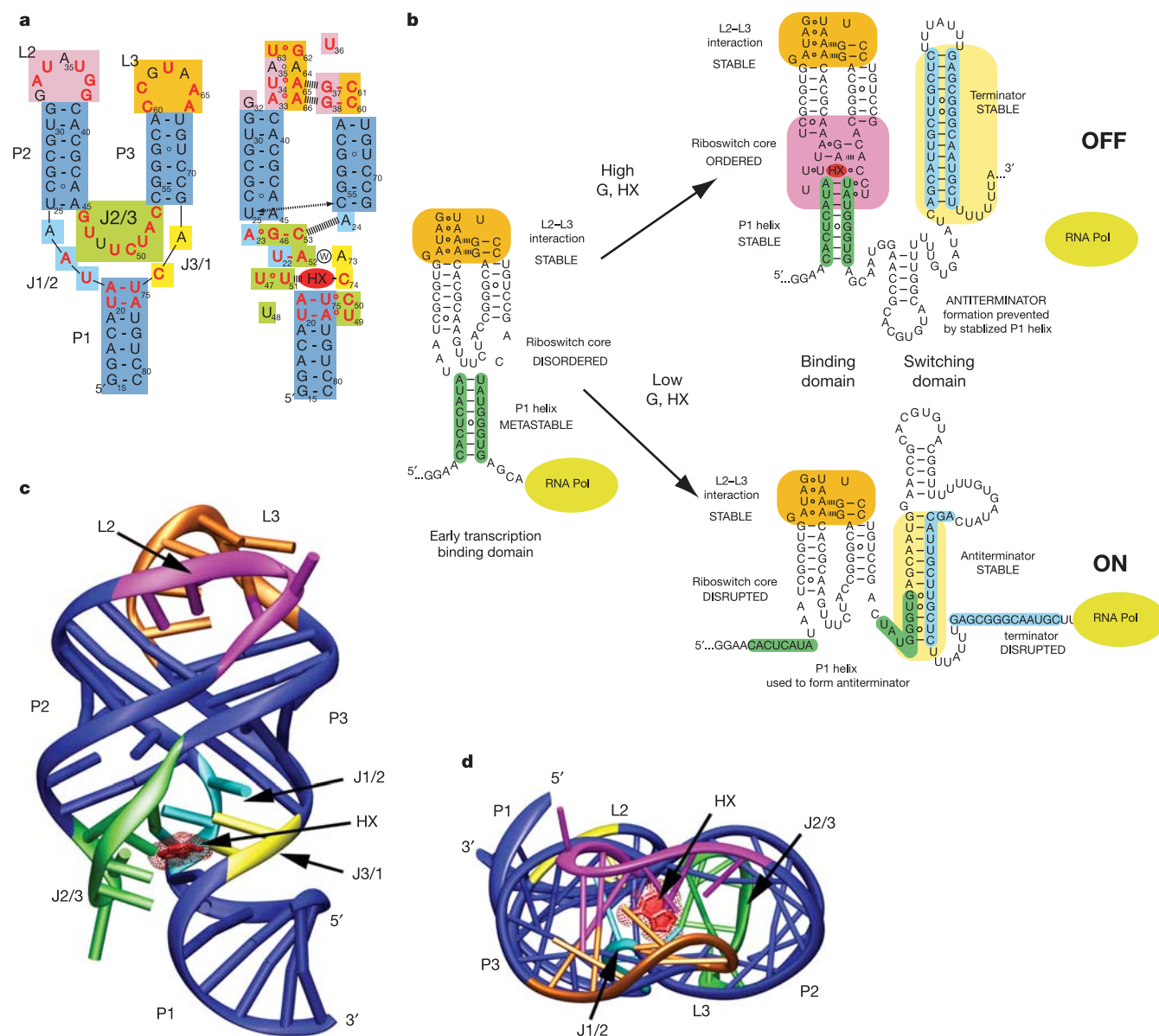


Figure 1 Secondary and tertiary structures of the guanine riboswitch–hypoxanthine complex. **a**, Left, secondary structure of the *xpt-pbuX* guanine-binding domain of the guanine riboswitch of *B. subtilis*⁵. Nucleotides conserved in more than 90% of known guanine riboswitches are shown in red; the numbering is consistent with that of the full-length mRNA. Coloured boxes correspond to structural features shown in Figs 2 and 3. Right, tertiary architecture of the hypoxanthine-bound form. Key tertiary interactions between the loops are shown as thick broken lines; a water-mediated triple is indicated by the circled 'w'. **b**, Gene repression by the guanine riboswitch in the 5' untranslated region

of mRNA. Initial transcription generates a binding domain that is primed to bind guanine (G) rapidly if it is at a sufficiently high concentration. Hypoxanthine (HX, top right) stabilizes the guanine-binding domain and particularly the P1 helix, forcing the mRNA to form a terminator element that halts transcription. In the absence of ligand (bottom right), an antiterminator forms, facilitating continued transcription. **c**, Ribbon representation of the three-dimensional structure of the RNA–hypoxanthine complex. The hypoxanthine ligand is shown in red, with its surface represented by dots. **d**, Top view of the complex, emphasizing the close packing of the P2 and P3 helices.

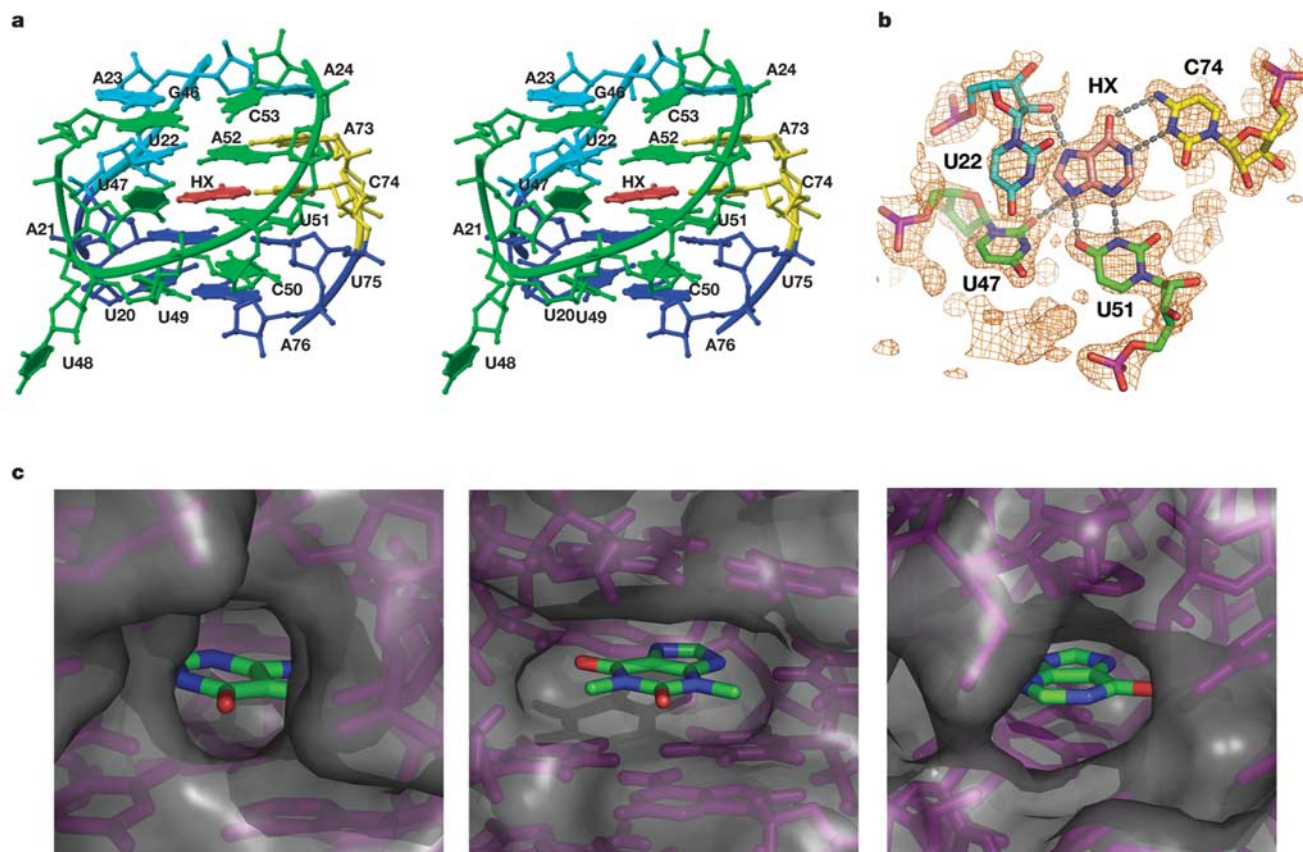


Figure 2 Recognition of hypoxanthine (HX) by the guanine-binding domain. **a**, Stereo view of the hypoxanthine-binding pocket in the three-way junction. **b**, Hydrogen-bonding interactions (grey broken lines) between hypoxanthine and the RNA. The final model (shown in stick representation) is superimposed on a simulated annealing $2F_o - F_c$ omit

map (orange cage), in which the atoms shown were excluded from the map calculation. **c**, Molecular surface representation of the binding pocket of the guanine riboswitch bound to hypoxanthine (left), compared with the theophylline-binding aptamer bound to theophylline⁹ (centre) and the *E. coli* purR repressor bound to hypoxanthine²⁸ (right).

through a single point mutation at nucleotide 74 from cytosine to uracil^{3,15}. Although the Watson–Crick pairing preference changes from guanine to adenine, the other interactions between the purine and the RNA are unchanged.

The tertiary architecture is stabilized through a unique loop–loop interaction capping helices P2 and P3 that is defined by two previously unobserved types of base quadruple. Each quadruple comprises a Watson–Crick pair with a noncanonical pair docked into its minor groove (G38–C60 and G37–C61 interacting with the A33–A66 and U34–A65 pairs, respectively; Fig. 3a). This arrangement bears a strong similarity to how adenosines pack into an A-form helix in the commonly found type I/II A-minor triple motif^{16,17}. Mutation of any one of these eight nucleotides would disrupt the intricate hydrogen-bonding network that cements the core of the loop–loop interaction, explaining their strict phylogenetic conservation. Stacked on these quadruples are two non-canonical pairs, including a side-by-side interaction between G62 and U63 akin to the A-platform motif¹⁸ and the bulged-G motif¹⁹

(Fig. 3b). The G62–U63 pair is further stabilized through hydrogen bonding to the backbone of the opposite loop.

The interaction between the two terminal loops is essential for ligand binding by the guanine riboswitch. Differences in the stability of the RNA in the absence and presence of guanine, as determined by in-line probing experiments³, indicate that this element of the RNA tertiary structure forms independently of guanine or hypoxanthine. Replacement of the wild-type loops with stable UUCG tetraloops²⁰ (Supplementary Fig. S3), which eliminates the tertiary interaction, abolishes the ability of the riboswitch to recognize hypoxanthine; this shows that it is crucial for promoting a high-affinity interaction (Fig. 4). Thus, although this tertiary interaction does not contact the ligand directly, it is significant in globally organizing the riboswitch for purine recognition. Similarly, natural sequences of various hammerhead ribozymes contain loop–loop interactions that significantly accelerate their rate of cleavage under physiological conditions^{21,22}. In each, the tertiary interaction constrains the RNA in a way that allows the three-way junction containing the functional site to respond to physiological concentrations of ligand or Mg^{2+} ions.

In high Mg^{2+} ion concentrations (20 mM $MgCl_2$), the guanine riboswitch has a very high affinity for guanine and hypoxanthine (an observed dissociation constant (K_d) of 5 nM and 50 nM, respectively)³. In *Escherichia coli*, however, repression of transcription of the *xpt-pbuX* operon by the purR repressor occurs in response to 1–10 μ M concentrations of purine. To determine whether the riboswitch responds to similar concentrations of purine, which probably reflect physiological levels, we determined its affinity for hypoxanthine by isothermal titration calorimetry at

Table 1 Thermodynamic parameters for hypoxanthine binding at 30 °C

$MgCl_2$ (mM)	K_d (μ M)*	ΔH_{obs} (kcal mol ⁻¹)*	ΔS_{obs} (cal mol ⁻¹ K ⁻¹)
20	0.732 $\pm$ 0.034	-33.5 $\pm$ 0.43	-82
5.0	1.34 $\pm$ 0.094	-28.4 $\pm$ 0.48	-67
1.0	2.99 $\pm$ 0.19	-35.4 $\pm$ 0.93	-91
0.25	4.00 $\pm$ 0.19	-41.9 $\pm$ 0.39	-113
0†	ND	ND	ND

*The reported errors represent the s.e.m. of the nonlinear least squares fit to the data. ND, no detectable binding.

†This reaction contained 2 mM Na_2-EDTA .

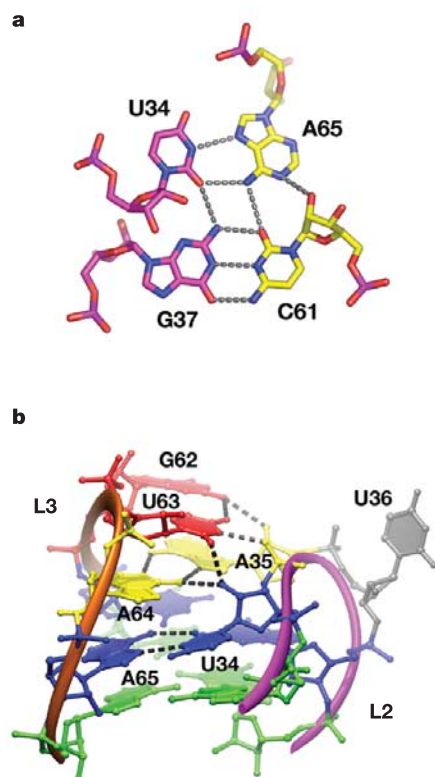


Figure 3 Stabilization of the tertiary architecture. **a**, One of two base quartets that form the core of the loop-loop contact. The carbon atoms are coloured as in Fig. 1. **b**, Side view of the loop-loop interaction, emphasizing the arrangement of base pairs and quartets. The bases of the quartet shown in **a** are coloured blue, with the hydrogen bonding between A65 and U34 shown for orientation; the bases of the other quartet are coloured green. The A35-A64 pair is shown in yellow, with hydrogen bonds emphasizing its interactions with the 2'-hydroxyl group of U34. The capping G62-U63 pair is shown in red.

varying ionic conditions (Table 1). At a more physiological ionic strength (0.25–1 mM Mg^{2+}), the RNA showed an affinity for hypoxanthine (observed $K_d = 3\text{--}4\text{ }\mu\text{M}$) similar to that of the purR repressor protein (observed $K_d = 9\text{ }\mu\text{M}$ for the *E. coli* variant²³). Thus, both RNA- and protein-based regulatory mechanisms seem to be tuned to respond to similar concentrations of intracellular metabolite. It has been also noted, however, that control of the related adenine riboswitch *in vivo* may be driven in part by the rate of ligand association with the riboswitch, complicating the relationship between intracellular ligand concentration and gene control⁴.

The structure further indicates how RNA directly transduces intracellular metabolite concentration into changes in gene expression through a proposed Rho-independent transcriptional regulation mechanism^{3,4}. Transcription of the initial ~90 nucleotides results in the formation of the guanine-binding domain, with the L2 and L3 loops interacting to begin to form the tertiary structure of the RNA (Fig. 1b). This partially organizes the three-way junction motif for efficient ligand binding, although the junction must be unstructured to some degree to allow access of the purine to the binding pocket. At sufficiently high concentrations of guanine or hypoxanthine, the nucleobase binds the pocket, stabilizing the short P1 helix through stacking interactions and base triples with J2/3 and preventing incorporation of P1 nucleotides into an antiterminator element. The mRNA can then form a classic Rho-independent terminator stem-loop, and transcription stops. By contrast, in low intracellular concentrations of guanine or hypoxanthine, the 3' side of the isolated P1 helix

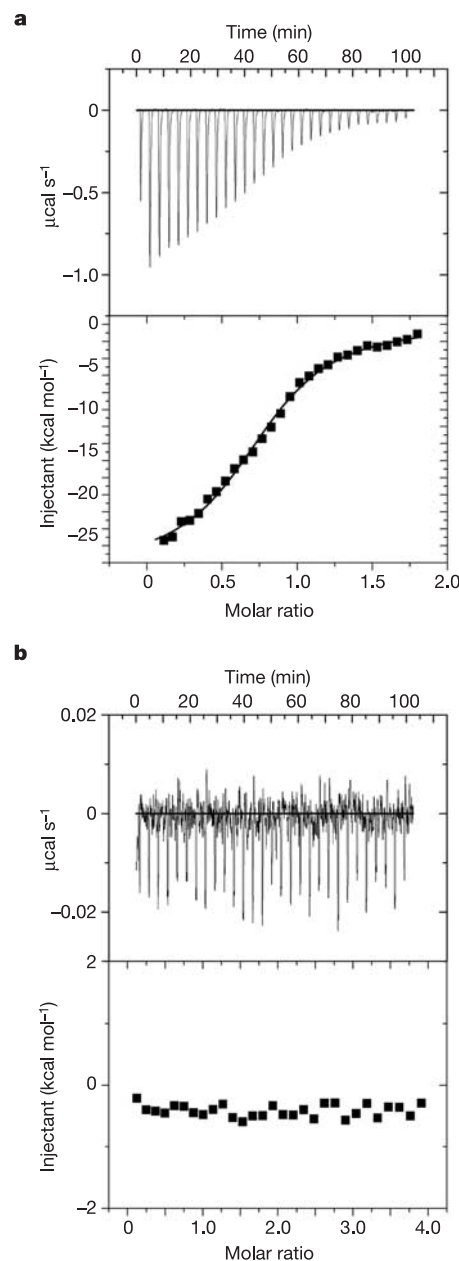


Figure 4 Estimation of the affinity of the riboswitch for hypoxanthine. Shown are the isothermal titration calorimetry curves for the wild-type guanine-binding domain (**a**) and for the guanine-binding domain lacking the tertiary interaction (**b**) with hypoxanthine at 30 °C in buffer containing 10 mM K^+ -HEPES (pH 7.5), 100 mM KCl and 5 mM $MgCl_2$.

is readily conscripted to form a stable antiterminator element, facilitating continued transcription. Thus, hypoxanthine is the keystone for the riboswitch, enabling the riboswitch to direct mRNA folding along two different pathways through its ability to stabilize one conformation over another, resulting in an effective biosensor of intracellular guanine, hypoxanthine and xanthine concentrations. □

Methods

Crystals

We synthesized and purified RNA by a native affinity-tag purification method²⁴ and exchanged it into a buffer containing 10 mM K^+ -HEPES (pH 7.5) and 1 mM hypoxanthine. Crystals were grown by mixing this solution in a 1:1 ratio with mother liquor (containing 25% PEG 3,000 (w/v), 200 mM ammonium acetate and 10 mM cobalt hexamine) and incubating it for 2–3 weeks at room temperature.

Data collection and processing

A single wavelength anomalous diffraction experiment was carried out at the CuK α wavelength on crystals cryoprotected in mother liquor plus 25% 2-methyl-2,4-pentandiol; clear diffraction was observed to at least 1.8 Å resolution. We indexed, integrated and scaled the data using D*TREK²⁵, identified heavy atom sites with SOLVE²⁶, and carried out refinement with CNS²⁷ to obtain a model with final R_{xtal} and R_{free} values of 17.8% and 22.8%, respectively. The model contains all RNA atoms except A82, for which electron density was not observed, along with the hypoxanthine ligand. For additional experimental details and references, see Supplementary Information.

Received 28 July; accepted 13 September 2004; doi:10.1038/nature03037.

- Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.* **20**, 44–50 (2004).
- Mandal, M. & Breaker, R. R. Gene regulation by riboswitches. *Nature Rev. Mol. Cell. Biol.* **5**, 451–463 (2004).
- Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113**, 577–586 (2003).
- Johansen, L. E., Nygaard, P., Lassen, C., Agerso, Y. & Saxild, H. H. Definition of a second *Bacillus subtilis* *pur* regulon comprising the *pur* and *xpt-pbuX* operons plus *pbuG*, *nupG* (*yxjA*), and *pbuE* (*ydhL*). *J. Bacteriol.* **185**, 5200–5209 (2003).
- Ellington, A. D. & Szostak, J. W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
- Gold, L., Polisky, B., Uhlenbeck, O. & Yarus, M. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763–797 (1995).
- Silverman, S. K. Rube Goldberg goes (ribo)nuclear? Molecular switches and sensors made from RNA. *RNA* **9**, 377–383 (2003).
- Seetharaman, S., Zivarts, M., Sudarsan, N. & Breaker, R. R. Immobilized RNA switches for the analysis of complex chemical and biological mixtures. *Nature Biotechnol.* **19**, 336–341 (2001).
- Zimmermann, G. R., Jenison, R. D., Wick, C. L., Simmorre, J.-P. & Pardi, A. Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. *Nature Struct. Biol.* **4**, 644–649 (1997).
- Fan, P., Suri, A. K., Fiala, R., Live, D. & Patel, D. J. Molecular recognition in the FMN–RNA aptamer complex. *J. Mol. Biol.* **258**, 480–500 (1996).
- Baugh, C., Grate, D. & Wilson, C. 2.8 Å crystal structure of the malachite green aptamer. *J. Mol. Biol.* **301**, 117–128 (2000).
- Koizumi, M., Soukup, G. A., Kerr, J. N. & Breaker, R. R. Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. *Nature Struct. Biol.* **6**, 1062–1071 (1999).
- Leulliot, N. & Varani, G. Current topics in RNA–protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**, 7947–7956 (2001).
- Williamson, J. R. Induced fit in RNA–protein recognition. *Nature Struct. Biol.* **7**, 834–837 (2000).
- Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Struct. Mol. Biol.* **11**, 29–35 (2004).
- Doherty, E. A., Batey, R. T., Masquida, B. & Doudna, J. A. A universal mode of helix packing in RNA. *Nature Struct. Biol.* **8**, 339–343 (2001).
- Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA* **98**, 4899–4903 (2001).
- Cate, J. H. *et al.* RNA tertiary structure mediation by adenosine platforms. *Science* **273**, 1696–1699 (1996).
- Correll, C. C., Beneken, J., Plantinga, M. J., Lubbers, M. & Chan, Y. L. The common and the distinctive features of the bulged-G motif based on a 1.04 Å resolution RNA structure. *Nucleic Acids Res.* **31**, 6806–6818 (2003).
- Molinaro, M. & Tinoco, I. Jr Use of ultra stable UNGC tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. *Nucleic Acids Res.* **23**, 3056–3063 (1995).
- De la Pena, M., Gago, S. & Flores, R. Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. *EMBO J.* **22**, 5561–5570 (2003).
- Khvorova, A., Lescoute, A., Westhof, E. & Jayasena, S. D. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nature Struct. Biol.* **10**, 708–712 (2003).

- Choi, K. Y. & Zalkin, H. Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *J. Bacteriol.* **174**, 6207–6214 (1992).
- Kieft, J. S. & Batey, R. T. A general method for rapid and nondenaturing purification of RNAs. *RNA* **10**, 988–995 (2004).
- Pflugrath, J. W. The finer things in X-ray diffraction data collection. *Acta Crystallogr. D* **55**, 1718–1725 (1999).
- Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron Radiat.* **11**, 49–52 (2004).
- Brunker, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by α helices. *Science* **266**, 763–770 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Edwards for maintaining and managing the Biochemistry Division X-ray Crystallography facility; and T. Cech, A. Pardi, D. Wuttke, J. Kieft and R. Rambo for discussions and comments on the manuscript. This work was funded in part from a grant from the Research Corporation and the University of Colorado Butcher Biotechnology Initiative. S.D.G. was supported in part by a NIH predoctoral training grant.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.T.B. (robert.batey@colorado.edu). The atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank under accession number 1U8D.

corrigendum

The genome of *Cryptosporidium hominis*

Ping Xu, Giovanni Widmer, Yingping Wang, Luiz S. Ozaki, Joao M. Alves, Myrna G. Serrano, Daniela Puiu, Patricio Manque, Donna Akiyoshi, Aaron J. Mackey, William R. Pearson, Paul H. Dear, Alan T. Bankier, Darrell L. Peterson, Mitchell S. Abrahamsen, Vivek Kapur, Saul Tzipori & Gregory A. Buck

Nature **431**, 1107–1112 (2004).

The GenBank accession number was supplied incorrectly as AAEL000000. The sentence should have read: ‘The sequences reported in this paper have been deposited in GenBank under the project accession number AAEL000000000’. In addition, the received date should have been 14 March 2004. □

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejp.com

naturejobs

Election returns

The victory of George W. Bush over his Democrat challenger John Kerry in the US presidential election this month suggests that the United States is becoming an increasingly divided country both politically and culturally. But it may also signal that the country is losing its status as a premier place to launch a scientific career. A week after the polls closed, a survey by 122 US graduate schools showed that enrolment by foreign graduate students across the board had dropped by 6% — the third consecutive decline. The slide in numbers began after the Bush administration tightened visa restrictions following the terrorist attack of 11 September 2001.

This decline may well continue — Bush's re-election is likely to reinforce perceptions that the United States is difficult to enter. This was evident from the recent survey, even though the Department of State has tried to streamline the visa application process and some US universities have taken a more active role in securing visas for prospective students.

In the sciences and engineering, foreign students make up about 50% of enrolment in the United States. Even before the Bush era, domestic enrolment in many of these fields had been declining. For the scientific job seeker in the United States, this could mean that it will be easier to get accepted at a top institution and get funded.

The potential long-term effects are harder to see, but are perhaps more significant. Intellectual property and spin-offs tend to come from the best research institutions, fuelled by bright young scientists. If this enrolment trend continues long-term, scientific jobs may shift to countries offering a more open visa policy — as long as they can also match the funding and infrastructure offered by the United States. And that would leave the United States playing second fiddle.

Paul Smaglik

Naturejobs editor



Contents

SPECIAL REPORT

Putting pen to paper p418

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Careers in journalism can be rewarding for scientists who have a way with words. Virginia Gewin reveals what it takes to be a scribe.

If you love science but hate lab work, trading the monotony of running gels for the creative buzz of writing may sound appealing. But before you abandon the bench and grab your laptop, make sure you understand the odds and options ahead. A staff position as a science journalist is as rare as it is coveted. In fact, many more opportunities exist for public information officers or freelance writers. Whatever the position, the ability to communicate is the common denominator.

People who write about science for a living fall into two broad categories. “There’s science journalism and science writing, and those are two different things,” says Deb Blum, president of the US National Association of Science Writers (NASW), which promotes both roles. Of the 2,300 members in the NASW, some 1,400 are journalists — only 500 of whom have staff positions. The remaining 900 members are science writers working as public information officers at universities, government science agencies or research institutions. These public relations (PR) positions are stable, well-paid jobs that can allow writers to hone their skills, and can still lead to journalistic pursuits.

THE JOURNALISTIC ANGLE

After years of a depressed market around the world, Blum says that journalism is undergoing a revival that reflects an upturn in US advertising dollars. But staff journalism jobs are, by nature, rare, and you can count on stiff competition for the few permanent positions at newspapers and magazines. “There may be 40 to 50 formal science correspondents for high-powered, high-ranking UK newspapers or television,” says Toby Murcott, a freelance science journalist in the United Kingdom. But a university will typically have more science writers than the relatively few magazines and newspapers that employ such specialists.

Although journalism is on an upswing in the United States, this effect has yet to cross the Atlantic. The global economic crisis continues to suppress advertising. “The biggest newspaper in Denmark had a very popular science section, but cut it away because there were no adverts in it,” says Jens Degett, director of the European Science Foundation’s Communication and Information Unit.

In Germany, editors solved the advertising crisis by sacking the staff. Now, 80% of the country’s science

journalists are freelancers, according to Hanns Neubert, vice-president of the umbrella organization the European Union of Science Journalists’ Associations.

The numbers don’t lie. Freelancers are the mainstay of science journalism. But Neubert and Blum agree that successful freelancers navigate the rocky waters by doing one thing — networking. “Network, network, network and then do some networking,” says Murcott. Blum cautions that the world of science journalism is very small. Much more than in many other professions, careers are built on who you know. The hurdle is getting to know the elusive editors.

In the United States, the would-be science journalist should consider attending the annual American Association for the Advancement of Science (AAAS) meeting. Editors typically trawl the February meeting for new writing talent, and the NASW piggybacks its annual conference with the AAAS meeting.

A number of editors and institutions conduct interviews for internships at the AAAS meeting. “I wasn’t aware that Fermilab offered internships before I went to the AAAS meeting,” says Davide Castelvecchi, an aspiring Italian science journalist and one of this year’s four science-writing interns at the accelerator lab in Batavia, Illinois.

TACKLING PUBLIC RELATIONS

Proving that writing skill is at the heart of both PR and journalism, Castelvecchi is using examples of his work at Fermilab to apply for jobs and internships in both sectors. The common denominator is communicating science in an accurate yet compelling way — an intangible skill that plays no part in obtaining a PhD.

Odds are that he will find the higher-paying job



“Don’t do this just because you can’t find something else, it is difficult to survive.”

Jens Degett, head of communication for the European Science Foundation.

Advice from the pros

“One thing you have to do the minute you walk into a science-writing programme is build your portfolio. You want to graduate as some kind of known quantity.” — Deb Blum, president, National Association of Science Writers

“I’d counsel reporters to do more reporting — flush out details that make a story sing.” — Steve Pelletier, editor, *HHMI Bulletin*

“Perhaps the most practical advice specific to freelancers is this: have a partner with a day job and benefits.” — Deb Blum



"People shouldn't make the mistake of thinking that becoming a science journalist is easier than becoming a scientist"
John Wilkes, director of the science-communication programme at the University of California, Santa Cruz.

with an institution or university. In the United States, the drive for media attention has reached unprecedented levels. So much so, according to Blum, that the bulk of science-writing jobs are now with institutions.

Although many point to a deep divide between science journalism and PR, Michel Claessens, head of communications for the European Commission's Research Directorate-General, notes that there are quite a few similarities. "Doing science PR at the commission is not fundamentally different from journalism," he says.

"We have to identify good stories, provide new messages and conduct interviews."

But high-profile magazines are increasingly being published by institutions. These outlets allow writers to display their talent for narrative — a quality any editor appreciates. As much as 75% of the content of magazines produced by institutions such as the Howard Hughes Medical Institute in Maryland and Fermilab is written by freelancers. In addition, these jobs can lead to special projects routinely farmed out to specialized writers. "We freelance a great number of projects including features on the website, some

news stories for the website, and special projects, such as reports or books," says *HHMI Bulletin* editor Steve Pelletier.

Europe is just beginning to attain the level of media relations that US institutions have honed. "Being media savvy is rather new over here," says Beate Kittl, science editor at Switzerland's *Facts* magazine. "It's really just happened at the government and university levels in the past five years." In fact, Kittl has trouble keeping good freelancers because they are attracted to the better-paid, steady work of the emerging PR positions.

TRAINING OPPORTUNITIES

There is no substitute for 'clips' — published proof of your abilities — in the world of science writing. But, here's the catch-22. "Editors are happy to publish you if you've already been published," says Blum.

"Editors are less willing than ever to take a chance on an unknown person," says John Wilkes, director of the science-communication programme at the University of California, Santa Cruz. "You almost have to come out of one of the best programmes to be eligible at the best internship sites."

Fortunately, advanced degree courses in science communication are beginning to crop up all over the world. In addition to the ten students accepted into Wilkes's programme each year, students will find MS courses at Boston University, Johns Hopkins University, and just recently the Massachusetts Institute of Technology. In Europe, Britain tends to lead the way, but such programmes exist in most European countries (see Web Links).

The Wellcome Trust in the United Kingdom, in conjunction with the Association of British Science Writers, funds seven bursaries for students interested in science journalism. The various science-writing associations are the best resources for professional development, often offering fellowships, short courses and seminars.

For those seeking a more direct route with hands-on experience, Wilkes suggests writing for a small, local newspaper in any way possible. Gail Cardew of the Royal Institution of Great Britain recommends that graduate students try contacting a science PR office, for example within a university, research council or science-communication organization. "They would then gain experience both in how the media works and, if their special interest was in writing, composing press releases," she says.

In the United States, summer fellowships and programmes can help you get a foot in the door. For example, the AAAS Mass Media Fellowship available to PhD students is a great way to make the right connections. "Over the past 30 years, almost half of our 450 fellows have remained in science journalism in some way," says Judy Kass, project director of the fellowship. "Given the immediate impact that cutting-edge science has on our lives today, it is increasingly important to communicate research findings to the public."

If there's a half-decent writer inside you fighting to get out, you will succeed. "There are a lot of people wanting to get into this peculiar business," says Murcott. "If you can do it, you will survive. The only way to find out is to try."

Virginia Gewin is a freelance science writer based in Portland, Oregon.

CORBIS

Web links

University of California, Santa Cruz
scicom.ucsc.edu
 Johns Hopkins University
www.jhu.edu/~writsem/sciwrit/index.html
 Boston University
www.bu.edu/com/jo/science/ms_txt.htm
 Massachusetts Institute of Technology
web.mit.edu/sciwrite
 AAAS Mass Media Fellow Program
ehrwab.aaas.org/massmedia.htm
 Santa Fe Science-Writing Workshop
sciwrite.org/sciwrite/sciwrite.html
 Association of British Science Writers bursaries
www.absw.org.uk/bursaries.htm
 Imperial College London
www.scicom.hu.ic.ac.uk/Welcome.html
 University of the West of England
www.uwe.ac.uk/fas/courses/environment/mscscicomm.html
 Higher Journalism School, Lille
www.esj-lille.fr
 Free University Berlin
www.kommwiss.fu-berlin.de/wissjour.html
 University of Milan
users.unimi.it/%7Espharm/comscient/index.htm
 International School for Advanced Studies, Trieste
mcs.sissa.it
 Umeå University
www.umu.se/kultmed/utbildning/vj/index.html